International Journal of Instruction *January 2021 ● Vol.14, No.1*
e-ISSN: 1308-1470 ● www.e-iji.net *p-ISSN: 1694-609X*
pp. 647-660

Article submission code: 20190819042301
Received: 19/08/2019
Revision: 18/07/2020
Accepted: 10/08/2020
OnlineFirst:10/11/2020

The Comparison of Item Test Characteristics Viewed from Classic and Modern Test Theory

Bambang Subali
Prof., Universitas Negeri Yogyakarta, Indonesia, *bambangsubali@uny.ac.id*

Kumaidi
Universitas Muhammadiyah Surakarta, Indonesia, *kum231@ums.ac.id*

Nonoh Siti Aminah
Universitas Sebelas Maret, Indonesia, *nonoh_nst@yahoo.com*

This research aims at comparing item characteristics of instruments for assessing the level of mastery in scientific method for elementary students as they were analyzed using Classical Test Theory (CTT) and Item Response Theory (IRT). The two analyses are usually done separately, for difference object, in this moment it was analyzed simultaneously for the same issue, that up two know it has not be done yet. This to ensure weather or not both the models of item analysis get the same result. The tests were developed in two different types namely a true-false and multiple choice. The multiple choice test consists of items with two and three options. Sample students from grade 1 to 6 were chosen consist of 234 school classes. The item responses were analyzed using a Quest Program. Results of the analysis show that item difficulty indexes for CTT and IRT are identical such as easy items on CTT are also identified as easy items on IRT, or vice versa. Based on CTT analysis results, three-option items are more difficult than the two-option items while IRT analysis results show that the three-option and two-option have similar difficulty indexes. The item difficulty indexes analyzed by CTT and IRT shows similar patterns. This research may reveal further information on the characteristics of test items which can contribute to the test development.

Keywords: assessment, classical test theory, item response theory, scientific methods, test theories

INTRODUCTION

The mastery of science process skills (SPS) has become one of the requirements for the teaching of Science. SPS in science learning should be introduced to learners as early as possible. Since SPS was arranged as scientific methods comprise of a series of complex scientific processes and its learning must be performed steps by steps and aspect by aspect, even sub-aspect by sub-aspect gradually. Thus, teaching scientific methods to

Citation: Subali, B., Kumaidi. & Aminah, N. S. (2021). The Comparison of Item Test Characteristics Viewed from Classic and Modern Test Theory. *International Journal of Instruction*, 14(1), 647-660. https://doi.org/10.29333/iji.2021.14139a

elementary students means teaching the students to master every aspect and sub-aspect of SPS (Subali, Kumaidi & Aminah, 2016).

In Indonesia, Ministry of Education and Culture of The Republic of Indonesia has emphasized the importance of SPS mastery for students since the implementation of School-Level Based Curriculum in 2006 and 2013 curriculum. In the 2013 curriculum, SPS should be applied in simple scientific methodology (SM) for science problem solving since grade IV.

To measure the mastery of skills in small scale in the classroom, criteria-based measurement is employed. Meanwhile, to measure the mastery of skills in large scale, norms-based measurement is used (Frisbie, 2005; Gronlund, 1998). Measurements in small scale can even be developed using alternative assessment as explored by Stears and Gopal (2010). The analysis of the outcomes shows that learners learn much more than the tests although what they learn is not necessarily science. The implication is that they explore the assessment of science concepts, as well as assessment of outcomes of science.

Studies of alignment measure the match, or the quality of the relationship, between a state's standards and its tests. That match can be improved by changing the standards, the tests, or both (Edwards, 2010). If it is carried out in a large scale, the norms-based measurement should be employed. Similarly, when SPS or SM is measured at the end of the program, or performed on a certain region, norms-based measurement is employed. This measurement is also performed when a supervisor wants to monitor and evaluate what teachers have taught to the students. In this case, the supervisor must use standardized tests in the form of confirmatory tests. It should be noted that monitoring SPS or SM mastery on a large scale (district or national level) using standardized tests provides a lot of benefits economically both between regions and time.

LITERATURE REVIEW

Scientific method is a prerequisite for obtaining scientific products (Carin & Sund, 1989). Brum & McKane stated on their book prepared by LeBoffe & Wisehart (1989) also argued that following systematic scientific processes, science process skills will become a scientific method. Through a series of scientific processes, scientific products in the form of new facts, concepts, and principles are obtained (Carin & Sund, 1989).

According to Bryce, McCall, MacGregor, Robertson and Weston (1990), the components of scientific methods include a number of basic and process skills that will produce investigative skills when they are compiled systematically. Meanwhile, Rezba (2007) use the term basic skills to refer as skills when they are integrated and will generate integrative skills in the form of scientific methods.

According to Bryce, et al. (1990), basic skill aspects consist of some sub-aspect skills namely (a) observing using senses, (b) recording data or information, (c) following instructions, (d) classifying, (e) measuring, (f) manipulating motions, (g) implementing procedures and using equipment, and (h) predicting. Process skills include sub-aspect skills of (a) inferencing, and (b) selecting procedures. But, Subali (2009) in his study indicates that the sub-aspect of predicting is more precisely classified in the aspects of processing. Moreover, Subali (2009) adds that skills of investigation include the sub-aspect skills of (a) planning investigation, (b) conducting investigation, and (c) reporting (in the written or spoken forms) the results of investigation.

The test for measuring student's mastery on scientific method can be in the form of skill written tests and performance tests (work sample tests). A good test development ideally refers to a learning continuum (NEWA, 2014), therefore it can measure the student skills. How to develop a written test has been widely presented by writers such as Millard (2012), Miller (2008), Popham (2005), Gronlund (1998), Gronlund and Linn (1990), Roid and Haladyna (1982). The kind of learning continuum of content knowledge in biology for senior high school students was formulated for example by Astuti and Subali (2017), Juniati and Subali (2017), Andriani and Subali (2017).

Acording to Frazier and Sterling (2009), if we want to assess to identify what student have learned and retained on their performance, we must employ a standardized test. With regard to this, the quality of the test must meet the requirements of validity and reliability, in addition to the various characteristics of the corresponding items, including the requirements of difficulty levels and differentiated items. When referring to criteria-based measurement, the quality of items can be analyzed using either Classical Test Theory (CTT) and Item Response Theory (IRT). The application of modern test theory for item analysis and constructing tests has been written by many authors, such as Hambleton and Jones (1993) and Le (2013). Hambleton and Jones (1993) review the application of CTT and IRT in analyzing and constructing tests. There are some shortcomings and advantages of using either theoretical model. The use of CTT has shortcoming as resulting item and person scores dependent but its applications is well known even by teachers. Those shortcomings in the use of CTT are resolved by the application of IRT, but the complexity of methodological payed its advantages.

According to Osteen (Le, 2013), in comparison to CTT, IRT is considered as the standard. Many testing programs still implement CTT in their design and assessment of test results. This is due to some advantages of CTT over IRT. For example, CTT describes the relationship between the true score and observed score in a linear fashion which makes CTT's models easy to understand and apply for many researchers. CTT offers smaller sample sizes than does IRT, CTT mathematical procedures are much simpler compared to IRT, parameter estimation in CTT is conceptually straightforward and requires minimum assumptions, making models useful and widely applicable, analysis do not require strict goodness of fit studies as in IRT. This comparison results are also given by Hambleton and Jones (1993).

Besides these advantages, CTT also has several disadvantages. One of them is that a testee score is a dependent test. It means that a test taker can get a higher score for easier tests and a lower score for difficult tests. With regard to these, there is no true value that can be obtained. It is impossible to be used as a basis for matching test items with students' ability level. In addition, an IRT advantage is that the children's ability levels and item difficulty levels can be plotted in a single line using a logit scale. Thus, the difficulty levels of the items can be compared with the ability of the test takers. Meanwhile, CTT item difficulty levels cannot be compared with students’ ability levels (Wright, 1999; Wright & Masters, 1982).

Researches on empirical item characteristics that attempt to compare CTT and IRT approaches has been widely used, such as by Pada, Kartowagiran and Subali (2016); Awopeju and Afolabi (2016); Petrillo, Cano, McLeod and Coon (2015), Zoghi and Valipour (2014); Qasem (2013); Adedoyin & Adedoyin (2013); Guler, Uyanik and Teker (2013); Stage (2003); Fan (1998). The comparisons of item characteristics according to CTT and IRT cover a wide range of applications, ranging from characteristic items such as difficulty and discrimination indexes, equating results using CTT and IRT approaches, reliability and the validity of the resulting scores. The results of comparison analyses using these two measurement theories show that in many respects CTT and IRT provide comparable results. This means that the results using CTT was not much different from the results of the analysis using IRT.

The focus of this research is developing the test to measure scientific method mastery because this measurement employs norms-based measurement which includes validating the test items using CTT and IRT approach. The problem is the extent to which the written test of scientific mastery methods influences the characteristics of the test type in terms of CTT and IRT. In this line, the problems are (a) whether the multiple choice test type with two options has different degree of difficulty compared to the multiple choice test with three options viewed from CTT and IRT, (b) whether the difficulty level of the analysis performed using CTT is the same as that of IRT, and (c) whether the items of a true-false test type have different characteristic viewed from CTT and IRT when the key answers are opposite (if a true-false model A, a statement is categorized as correct then in the B model the statement is changed to the false category).

METHODS

The scientific methods test was constructed after a learning continuum of the scientific aspects method had been prepared by a team of teachers and teacher educators (Subali, 2009). Also, it referred to the learning continuum of SPS developed by Subali (2009) which had been revised by Subali and Mariyam (2013) and validated by eight educational experts judgment from the Sebelas Maret University and Yogyakarta State University.

The test sets which had been prepared were administered to samples of elementary students in two Technical Management Units (TMU) or the smallest unit of school groups in Yogyakarta City and one TMU in Sleman Regency, namely TMU of Kalasan. Among those three TMUs, each was taken 13 elementary schools as samples based on the assessment of the supervisors. Each School was taken one class each from the first grades up to the sixth grades. If schools had parallel classes, all the parallel classes in the schools were sampled.

The sample consists of 78 classes ranging from the first up to the sixth grades from each TMU taken from three TMUs in Yogyakarta Province. Therefore, the total of sample is 234 classes of which the class size ranges from 20 to 30 students. This number of samples is also expected to meet the requirement for testing which are analyzed using Graded Model in which the minimum number of test takers required for research is 250 test takers (Muraki & Bock, 1998). In addition, the item analysis is performed using the Quest Program (Adam & Kho, 1996).

The tests were developed in two different types namely a true-false and multiple choice items. The multiple choice test (MCT) form consists of MCT items with two options (MCT2O) and MCT items with three options (MTC3O). For a true-false test (TFT), there are two models, TFT set A (TFTA) and TFT set B (TFTB) of which the key answers are opposite. The test item development considered the aspects of substance, construction, and language. The same test set was given to all grades (grade 1 until 6).) The number of test takers is different because 8 test set was administered simultaneously in each class and school therefore it depended on the number students in each class and school. For this reason, the number of test takers for each test set is different one another. They are presented in Table 1.

Table 1
Types of test set on scientific methods mastery along with the specification of questions, sub-aspects, scientific methods, and objects of natural objects

Type of test set	Type of question	sub-aspect groups of scientific methods	the specification of items related to the objects	Number of testee	Fit item with Rasch model	Reliability (Error of measurement/Internal consistency)
Test set 1	MCT2O	Grouping Code I	20 items (item number 1 to 20) are related to living thing objects	806	All items fit	0.72/0.72
		Grouping Code III	15 items (number 21 to 35) are related to non-living thing objects			
Test set 2	MCT3O	Grouping Code I	20 items (item number 1 to 20) are related to living thing objects	802	All items fit	0.70/0.70
		Grouping Code III	15 items (number 21 to 35) are related to non-living thing objects			
Test set 3	TFTA	Grouping Code I	20 items (item number 1 to 20) are related to living thing objects	740	All items fit	0.51/0.49
		Grouping Code III	15 items (item number 21 to 35) are related to non-living thing objects			
Test set 4	TFTB	Grouping Code I	20 items (item number 1 to 20) are related to living thing objects	758	All items fit	0.59/0.57
		Grouping Code III	15 items (item number 21 to 35) are related to non-living thing objects			
Test set 5	MCT2O	Grouping Code II	20 items (item number 1 to 20) are related to living thing objects	752	All items fit	0.70/0.72
		Grouping	15 items (item number 21			

		Code IV	to 35) are related to non-living thing objects			
Test set 6	MCT3O	Grouping Code II	20 items (item number 1 to 20) are related to living thing objects	730	All items fit	0.71/0.71
		Grouping Code IV	15 items (item number 21 to 35) are related to non-living thing objects			
Test set 7	TFTA	Grouping Code II	20 items (item number 1 to 20) are related to living thing objects	703	All items fit	0.54/0.51
		Grouping Code IV	15 items (item number 21 to 35) are related to non-living thing objects			
Test set 8	TFTB	Grouping Code II	20 items (item number 1 to 20) are related to living thing objects	707	All items fit	0.55/0.53
		Grouping Code IV	15 items (item number 21 to 35)are related to non-living thing objects			

Notes: 1) All items are in between the range of 0.77-1.30 of INFITMNSQ. It means they are fitted with Rasch model.

2) True false items of B model: the key answer is opposite to that of A model

Referring to Table 1 and the analysis of Rasch model, it is found that all items of the tests fit with the model; therefore all existing tests are declared to be “valid” instruments for measuring the students’ ability to think about the scientific methods. It is also found that by referring to the degree of the reliability coefficients, MCT2O and MCT3O are producing more reliable scores than TFTA and TFTB.

RESULTS

The comparison of item difficulty indexes (DI) as estimated by the CTT and the 1PL-IRT (also known as Rasch model) to the all types of test items namely MCT2O, MCT3O, and TFTA and TFTB are presented in Table 2.

Table 2
The comparison of DI of CTT and IRT results between MCT2O and MCT3O for test set with scientific methods sub-aspect codes of I-III and II-IV done by grade 1 to 3 and grade 4 - 6.

Scientific Sub-aspect Codes	Group of testee	Type of DI	Number of items MCT3O which are more difficult than items MCT2O	Total of items	%	Note
I-III	Grades 1-3 (N 397 for MCT2O and N 395 for MCT3O)	CTT DI	24	35	68.57	More than 50%
		IRT DI	19	35	54.29	More than 50%
	Grades 4-6 (N 409 for MCT2O and N 407 for MCT3O)	CTT DI	22	35	62.86	More than 50%
		IRT DI	16	35	45.71	Less than 50%
II-IV	Grades 1-3 (N 373 for MCT2O and N 363 for MCT3O)	CTT DI	24	35	68.57	Less than 50%
		IRT DI	13	35	37.14	Less than 50%
	Grades 4-6 (N 381 for MCT2O and N 367 for MCT3O)	CTT DI	25	35	71.43	Less than 50%
		IRT DI	17	35	48.57	Less than 50%
I-III	Grades 1-3 (N 397 for MCT2O and N 395 for MCT3O)	CTT DI	24	35	68.57	More than 50%
		IRT DI	19	35	54.29	More than 50%
	Grades 4-6 (N 409 for MCT2O and N 407 for MCT3O)	CTT DI	22	35	62.86	More than 50%
		IRT DI	16	35	45.71	Less than 50%
II-IV	Grades 1-3 (N 373 for MCT2O and N 363 for MCT3O)	CTT DI	24	35	68.57	Less than 50%
		IRT DI	13	35	37.14	Less than 50%
	Grades 4-6 (N 381 for MCT2O and N 367 for MCT3O)	CTT DI	25	35	71.43	Less than 50%
		IRT DI	17	35	48.57	Less than 50%

Table 2 shows that, based on the results of CTT analysis, many items are more difficult if they are arranged in the form of MCT3O than MCT2O. The number of items that is more difficult if tested using MCT3O than that of MCT2O is more than 50%, ranging from almost 63% to about 71%. This is found in all sub-aspect codes of scientific method, code I-III and code II-IV tested in the first grade up to the third one and the fourth grade up to the sixth one. Meanwhile, the analysis using IRT approach shows the opposite results. This occurs both for tests with scientific method sub-aspects of code I-III and code of II-IV. Whereas, the IRT analysis shows an increase in which the number of items that are more difficult using MCT3O than MCT2O is less than 50%. In addition, a test for the sub-aspect group of scientific method code I-III of the test takers group of the first up to the third grades is the only test of with DI is about 54%.

The following is presented the comparison of item DI both for items of MCT3O and MCT2O performed by testee groups of the 1st to 3rd grades or by the 4th up to 6th grades using test sets for the sub-aspect of the scientific method of code I-III and code II-IV. The results of CTT and IRT analysis are described in Figures 1-8 (see the appendix).

Figure 1 is presenting the comparison of item DI estimated by CTT and IRT on scientific methods mastery tests for sub-aspects of scientific method code I-III. The comparison of item DI based on test results performed by test takers of the first up to third grades for MCT2O is presented in Figure 1 and for MCT3O is presented in Figure 2. The comparison of item DI based on the test results performed by testee of the forth to sixth grades for MCT2O is described in Figure 3 and for MCT3O is described in Figure 4.

The next figures describe the comparison of item DI utilizing CTT and IRT approaches for the scientific method mastery of scientific method sub-aspects of code II-IV. The comparisons are based on the results of tests performed by testee of the first up to third grades. In this case, MCT2O is presented in Figure 5 while MCT3O is in Figure 6. Moreover, the next comparisons are based on test results performed by testee of the forth up to sixth grades. In this case, MCT2O is presented in Figure 7 while MCT3O is presented in Figure 8.

Figures 1 to 8 show that the results of CTT and IRT analysis have similar patterns, both for the sub-aspects of the scientific method of code I-III and II-IV, and either the testee group of the first up to third grades or the forth up to sixth grades. When an item analyzed using CTT is categorized as difficult, the results of IRT analysis are also categorized as difficult. Similarly, if an item analyzed using CTT is classified as an easy item, the result of IRT analysis is also classified as an easy item.

These results suggest that in term of small scale testing, such as in a classroom situation, the application of CTT as compared to the use of 1PL IRT (in this case also namely as Rasch model) might be justified. This suggestion supports the prior claim that classroom teachers may be more familiar and capable in using CTT than that of IRT due to its wide used in schools and easier of usage as presented by many studies. This finding may make teachers continually use CTT in their test development than that of 1 PL IRT which many teachers do not familiar.

The comparison of item DI of TFT model is presented in a reverse pattern. If a statement is true on TFTA, the statement is made false on TFTB. The results of CTT analysis compared with IRT are described in Figures 9-16 (see the appendix).

First, it is presented the comparison of the item DI analyzed using the CTT and IRT approaches for scientific method mastery of a set of TFTA of the scientific method sub-aspect code I-III (described in Figure 9) and TFTB (presented in Figure 10) for test takers of the first up to third grades of elementary school students (a test sample of TFTA consists of 350 persons and TFTB is 382 persons).

Second, it is presented the comparison of the item DI estimated by the CTT and IRT approaches for scientific method mastery of a set of TFTA items of the scientific method sub-aspect code I-III (described in Figure 11) and TFTB (presented in Figure 12) for testee of forth up to sixth grades of elementary school students (a test sample of TFTA consists of 390 persons and TFTB is 376 persons).

The next figure describes the comparisons of the item DI analyzed using CTT and IRT approaches for scientific method mastery of a set of TFTA of the scientific method sub-aspect code II-IV (described in Figure 13) and TFTB (presented in Figure 14) after being tested to test takers of the first up to third grades of elementary school students.

The next figure explains the comparisons of the item DI analyzed by CTT and IRT approaches for scientific method mastery of a set of TFTA of the scientific method sub-aspect code II-IV of a test set of A and a test set of B after being tested to test takers of forth up to sixth grades of elementary school students.

Figure 9 to 16 indicate the similar pattern of item DI analyzed by CTT and IRT for a test set A and B, both for the sub-aspects of the scientific method of code I-III and II-IV, and either answered by the test takers of the first up to third grades or test takers of the forth up to sixth grades. When an item of a test set A analyzed using CTT is more difficult than a test set B, the results of IRT analysis indicate the same findings as that of CTT. Similarly, if an item of a test set A analyzed using CTT is easier than a test set B, the results of IRT analysis indicate the same as that of CTT.

The following is presented the comparisons of item DI in which the test results from the first up to third graders are compared to that of the forth up to sixth graders. It is expected that the degree of difficulty decreases when the tests are administered by the higher grades.

Table 3
The comparison of DI of CTT and IRT analysis results between the items administered to the first up to third graders and the forth up to sixth graders for the test set with the sub-aspect of the scientific method of code I-III and II-IV

Sub-aspect codes of scientific methods	Test type	Type of difficulty index	Number of items tested to grade 1 - 3 which are more difficult for grade 4 – 6	Number of items	%	Notes
I-III	MCT2O	CTT DI	28	35	80.00	Far over 50%
		IRT DI	20	35	57.14	arround 50%
	MCT3O	CTT DI	29	35	82.86	Far over 50%
		IRT DI	18	35	51.43	arround 50%
	TFTA	CTT DI	28	35	80.00	Far over 50%
		IRT DI	17	35	48.57	arround 50%
	TFTB	CTT DI	28	35	80.00	Far over 50%
		IRT DI	16	35	45.71	arround 50%
II-IV	MCT2O	CTT DI	32	35	91.43	Far over 50%
		IRT DI	18	35	51.43	arround 50%
	MCT3O	CTT DI	30	35	85.71	Far over 50%
		IRT DI	18	35	51.43	arround 50%
	TFTA	CTT DI	27	35	77.14	Far over 50%
		IRT DI	19	35	54.29	arround 50%
	TFTB	CTT DI	29	35	82.86	Far over 50%
		IRT DI	18	35	51.43	arround 50%

Table 3 shows that the results of the analysis using IRT are more balance the CTT, regarding the items difficulty indexes which increase and decrease when associated with the group of test takers answering the tests.

DISCUSSION

***Figure 1-16 show a consistent result regarding the characteristics of items based on CTT and IRT for all test types, if the analysis is performed for the same test. The first findings show that the CTT analysis for most MCT3O items is more difficult than MCT2O. However, this is not found in IRT analysis results. The findings are valid for a test set of sub-codes I-III and II-IV of scientific methods tested to grade 1 to 3 and 4 to 6. The research conducted by Haladyna and Downing (1993) shows that MCT items rarely contain more than three useful options. Consequently, testing program personnel and classroom teachers may be better served by using MCT 2-or 3-option items instead of the typically recommended 4- or 5-option items. With regard to this, the number of choices will affect the quality of the test.

The second finding indicates that the item DI based on CTT and IRT shows the similar pattern for both MCT2O, MCT3O, and TFT type. If an item on MCT2O set belongs to be more difficult category according to the CTT similar results shows by the IRT. This is found either on the tests for the sub-aspects of the I & III or II & IV of scientific methods whether for testee of the first up to third graders or the forth up to sixth graders. This is also valid for MCT3O and TFT test type. If an item analyzed using CTT on TFTA is more difficult than TFTB, the results of the IRT analysis will be identical or opposite one.

This finding is relevant with the research results of Adedoyin and Adedoyin (2013); Awopeju and Afolabi (2016), Fan (1998); Petrillo et.al (2015); Zoghi and Valipour (2014). Fan (1998) conducted a research on the comparison of item response theory and classical test theory: an empirical comparison of their item/person statistics. The major findings include: (a) the person statistics (examinee ability estimates) from CTT were highly comparable with those from IRT for all three IRT models, (b) the item difficulty indexes from CTT were very comparable with those from all IRT models and especially from the Rasch model, (c) compared with item DI, the item discrimination indexes from CTT were somewhat less comparable with those from IRT, (d) both CTT and IRT item DI exhibited very high invariant results across samples, even across samples that were quite different (samples from high- and low-ability groups), (e) both the CTT and IRT item discrimination estimates were somewhat less invariant than their item difficulty estimates. The degree of invariant and consistency results in item discrimination based on CTT and IRT analyses were highly comparable.

Zoghi and Valipour (2014) compare CTT and IRT in estimating test item parameters in a linguistics test. This study was an attempt to assess the comparability test items parameter estimates between CTT and IRT models. To estimate the test items parameters in terms of item difficulty, item discrimination, and the responses given by the students to each item, CTT and IRT (2PL) models were used. Results suggested that CTT and IRT test items parameters are comparable.

The results of this study are also in line with the results of research performed by Adedoyin and Adedoyin (2013). Adedoyin and Adedoyin (2013) that use the Junior Certificate of mathematics test in Botswana conclude that the estimation results of the item DI categorized as an easy item based on CTT will also be identified as an easy item according to IRT (3PL), or vice versa. This suggests that the results are quite similar to results of this study. The similar findings are also presented by the research of Awopeju and Afolabi (2016), and Petrillo et. al (2015).

The results of the current study were in agreeable with those similar studies as discussed above. The implications of the results suggest that in term of small scale testing, such as in a classroom assessment, the usage of CTT may be recommended as it is simpler in analysis and easier to use by teachers. Beside that classroom, teachers may be more familiar and capable in using CTT than that of IRT due to its wide used and easier of usage in schools. This finding may support teachers continually use CTT in their test development than that of 1-PL IRT which many teachers do not familiar.

The third finding, which is specific to this study, shows that the results of CTT analysis of most items administered by the test takers of the first up to third graders are more difficult than those performed by the forth up to sixth graders. However, it is not found in IRT analysis results. The findings are valid for the test sets with sub-codes of scientific method of codes I-III and II-IV for MCT2O, MCT3O, and TFT types.

Refererring to the results of CTT, the principle above can be demonstrated, but the results are just the same when being analyzed using the IRT. With regard to this, it is necessary to examine further how exactly the ability of students who administer the tests. Based on the national curriculum, the scientific method is part of the process of science that must be taught. This means that students should be accustomed to practice using aspects of scientific methods. Therefore, the mastery of the scientific method aspects of students in the fourth up to sixth grades is higher than that of the first up to third graders. Based on the study conducted by Subali et al. (2016), teachers argue that aspects of the scientific method should have been taught in elementary school students.

CONCLUSION AND SUGGESTION

Based on the results and discussions, it can be concluded that the CTT and IRT analyses for item DI show similar or identical results in which easier items on CTT are also identified as easier items on IRT, or vice versa. In addition, other results show that most items of MCT3O types are more difficult than MCT2O ones. However, it is not found in the IRT analysis results. The item DI based on the CTT and IRT show similar patterns both for MCT2O, MCT3O, and TFT type. When MCT2O items are analyzed using CTT is included in a more difficult category, the IRT analysis will show the same results as the CTT. In addition, CTT analysis of most items administered to test takers of the first up to third graders is more difficult than the items administered to the forth up to sixth graders. This is not always found in the IRT.

It should be noted that IRT model is understood to be better to reveal the score of the test than CTT model because the result or the IRT model is not affected by characteristic group of test takers. Thus, from this result it suggest that test items can be analyzed using only CTT or IRT model separately depend on the practitioners. However, it is recommended to run both models simultaneously for other test items to get more data and make generalization.

ACKNOWLEDGEMENTS

The researchers would like to thank to the Directorate of Research and Community Service of the Ministry of Research, Technology and Higher Education which has provided a sponsor so that this research could be conducted.

REFERENCES

Adams, R.J., & Kho, S. (1996). *Acer quest version 2.1*. Camberwell, Victoria: The Australian Council for Educational Research (ACER).

Adedoyin, O.O., & Adedoyin, J.A. (2013). Assessing the comparability between classical test theory (CTT) and item response theory (IRT) models in estimating test item parameters. *Herald Journal of Education and General Studies*, 2(3), 107-114.

Andriani, A.E., & Subali, B. (2017). Teachers' opinion about learning continuum based on student's level of competence and specific pedagogical material in classification topics. *AIP Conference Proceedings 1868, 100001 (2017*). https://doi.org/10.1063/1.4995211

Astuti, L.D. & Subali, B. (2017). Teacher's opinions about learning continuum based on the student's level of competence and specific pedagogical materials on anatomical aspects. *AIP Conference Proceedings 1868, 100005 (2017).* https://doi.org/10.1063/1.4995215

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,* 12(28), 263-284.

Bryce, T.G.K., McCall, J., MacGregor, J., Robertson I.J., & Weston, R.A.J., (1990). *Techniques for assessing process skills in practical science: Teacher's guide*. Oxford: Heinemann Educational Books.

Carin, A.A., & Sund, R.B., (1989). *Teaching science through discovery*. Columbus: Merrill Publishing Company.

Edwards, N. (2010). An analysis of the alignment of the Grade 12 Physical Sciences examination and the core curriculum in South Africa. *South African Journal of Education*, 30, 571-590.

Fan, X. (1998). Item response theory and classical test theory: an empirical comparison of their item/person statistics. *Educational and Psychological Measurement,* 58(3), 357.

Frazier, W.M. & Sterling, D.R. (2009). Helping new science teacher: A how-to guide for experienced teachers. *The Science Teacher,* summer 2009: 34-39.

Frisbie, D.A. (2005). Presidential Address. Measurement 101l: Some Fundamentals Revisited. *Educational Measurement: Issues and Practice,* Fall:21-28.

Gronlund, N.E. (1998). *Assessment of student achievement* (9th ed). Boston: Allyn and Bacon.

Gronlund, N.E. & Linn, R.L., (1990). *Measurement and evaluation in teaching* (6th ed). New York: MacMillan Publishing company.

Guler, N., Uyanik, G.K., & Teker, G.T., (2013). Comparison of classical test theory and item response theory in terms of items parameters. *European Journal of Research on Education,* 2(1), 1-6.

Haladyna, T. M. & Downing, S. M. (1993). How many options is enough for a multiple-choice test item?. *Educational and Psychological Measurement*, 53(4), 999-1010.

Hambleton, R.K. & Jones, R.W. (1993). Comparison of classical test theory and item response theory and their applications to test development. An NCME Instructional Module. *Educational Measurement: Issues and Practice,* 12(3), 253-262.

Juniati, E. & Subali, B., (2017). Teacher's opinion about learning continuum of genetics based on student's level of competence. *AIP Conference Proceedings 1868, 100002 (2017).* https://doi.org/10.1063/1.4995212

Le, D. (2013). Applying item response theory modeling in educational research. Doctoral Dissertation. Ames St: Iowa State University. Available at http://lib.dr.iastate.edu/etd. Accessed 17 October 2015.

LeBofee, M. & Wisehart, G. (1989). *Study guide Biology: Exploring life* (prepared by LeBoffe, M. & Wisehart G). New York: John Wiley & Sons.

Millard, S. (2012). *Writing multiple choice and true/false exam questions. A good practice guide*. Kawili: Department of Pharmacy Practice, University of Hawai'I At Hilo.

Miller, P.W. (2008). *Measurement and teaching*. Munster: Patric W. Miller & Associates.

Muraki, E. & Bock, R.D. (1998). PARSCALE: IRT item analysis and test scoring for rating-scale data. Chicago: Scientific Software International, Inc.

NWEA (2014). *Measures of academic progress learning continuum*. Everett St: Northwest Evaluation Association. Portland, OR. Retrieved 17 October, 2015 from https://www.nwea.org/content/uploads/2015/09/MAP-Learning-Continuum-Brochure-JUL16.pdf.

Pada, A.U.T., Kartowagiran, B., & Subali, B. (2016). A separation index and fit items of creative thinking skills assessment. *Research and Evaluation in Education,* 2, 1-12.

Petrillo, J., Cano, S.J., McLeod, L.D., & Coon CD. (2015). Using classical test theory, item response theory, and Rasch measurement theory to evaluate patient-reported outcome measures: a comparison of worked examples. *Value in Health*, 18(1), 25-34. doi: 10.1016/j.jval.2014.10.005.

Popham, W.J. (2005). *Classroom assessment: What teachers need to know* (4th ed). Boston: Pearson Education, Inc.

Rezba, R.J., Sparague, C.S., Fiel, R.L., Funk, H.J., Okey, J.R., & Jaus, H.H. (2007). *Learning and assessing science process skills* (3rd ed). Iowa: Kendall/Hunt Publishing Company.

Roid, G.H., & Haladyna, Th.M. (1982). *A technology for test-item writing*. Oriando: Academic Press, Inc.

Stears, M., & Gopal, N. (2010). Exploring alternative assessment strategies in science classrooms. *South African Journal of Education*, 30, 591-604.

Subali, B., Kumaidi & Aminah, N.S. (2016). *Teacher's opinion about the learning and assessment of scientific method aspects and subaspects in natural science on elementary schools*. Paper presented at the International Conference on Education, Malang, 24 November.

Subali, B. & Mariyam, S. (2013). Pengembangan kreativitas keterampilan prcses sains dalam aspek kehidupan organisme pada mata pelajaran IPA SD. *Cakrawala Pendidikan*, 3.

Subali, B. (2009). *Pengukuran keterampilan proses sains pola divergen mata pelajaran biologi SMA di Provinsi DIY dan Jawa Tenga* [The measurement of science process skills of divergent pattern on Biology Subject at SMA in DIY and Central Java Province]. Doctoral Dissertation. Yogyakarta: Yogyakarta State University

Stage, C. (2003). Classical test theory or item response theory: the swedish experience, EM No 42, 2003, ISSN 1100-696X, ISRN UM-PED-EM--42—SE

Wright, B.D. (1999). Rasch measurement model. In GN Masters & JP Keeves. *Advances in measurement in educational research and assessment.* Amasterdam: Pergamon, An imprint of Elsevier Science.

Wright, B.D. & Masters, G.N. (1982). *Rating scale analysis*. Chicago: Mesa Press.

Zoghi, M. & Valipour, V. (2014). A comparative study of classical test theory and item response theory in estimating test item parameters in a linguistics test. *Indian Journal of Fundamental and Applied Life Sciences*, 4, 424-435.