JPII 8 (1) (2019) 39-51

Jurnal Pendidikan IPA Indonesia

http://journal.unnes.ac.id/index.php/jpii

# STUDENT ACHIEVEMENT BASED ON THE USE OF SCIENTIFIC METHOD IN THE NATURAL SCIENCE SUBJECT IN ELEMENTARY SCHOOL

**B. Subali[1], Kumaidi[2], N. S. Aminah[3], B. Sumintono[4]**

[1]Biology Education Department, Faculty of Mathematics and Natural Sciences, Universitas Negeri Yogyakarta, Indonesia
[2]Universitas Muhammadiyah Surakarta, Indonesia
[3]Universitas Sebelas Maret, Indonesia
[4]University of Malaya, Malaysia

**DOI: 10.15294/jpii.v8i1.16010**

Accepted: November 6th, 2018. Approved: March 25th, 2019. Published: March 28th, 2019

## ABSTRACT

This research aims at investigating elementary school student achievement based on the use ofscientific method in teaching science from the test item types, as reflected by the item difficulty indexusing the classical test theory (CTT) and modern test theory (IRT). The first stage in developing the test was preparing the learning continuum of scientific method aspects by referring to the learning continuum of science process skill as developed by the previous existing research. In this research, the learning continuum was validated by expert judgment. As the tests were administered/carried out at the same time, four sets of tests were developed and administered to students of Grade 1 to 6 in Yogyakarta and Sleman Regency in the 2016-2017 school year. Samples were taken from three Technical Management Units (TMUs). Three TMUs were determined by observing the distribution of school locations from the center to the suburbs. The items were analyzed using CTT and IRT.The results of the research show that the student achievement reflected by item difficulty index based on CTT and IRT indicates the same level of category except for several sub-aspects. Those items from certain testsindicate higher difficulty level for Grade 4 to 6 students than for Grade 1 to 3 students. This case is not relevant to the expected learning outcomes.



Keywords: scientific method, test types, CTT, IRT

## INTRODUCTION

An inquiry has become the heart of science education reform. Despite its prevalence in school and policy rhetoric, the term remains quite ambiguous. Generally, an inquiry becomes conflated with the scientific method, taught as a series of steps ranging from asking a question to drawing conclusions (Tang et al., 2010). According to Zeidan & Jayosi (2015), science education is aimed at teaching students how to get involved in an inquiry. Science process skills are a necessary tool to produce and use scientific information, to perform scientific research, and solve problems. The integration of explicit, reflective instruction about the nature of science (NOS) and scientific inquiry (SI) in traditional science content is addressed as a means through which the development of scientific literacy is fostered (Lederman et al., 2013).

For science instruction to be productive, teachers should consider physical and mental skills to generate and test reliable knowledge and generalization. In learning science, the process

***Correspondence Address**
E-mail: bambangsubali@uny.ac.id

skills include observing, measuring, classifying, inferring, predicting, hypothesizing, identifying variables, experimenting and interpreting data. Generally, these are processes carried out by scientists during investigations.The aim of this approach is for pupils to learn and obtain an understanding through the development of their own ideas. Therefore, the teaching method that should be adopted for the teaching of science must be carefully considered and should encourage the development of science process skills by the learners. The attention of many science educators has continued to be directed at searching for such appropriate methods of science teaching (Nweke et al., 2014). In Biology learning, high-school biology students meet authentic science regarding their participation in a science outreach programme (Tsybulsky et al., 2019).

Science Learning depends on the kinds of strategies used by teachers. For example, project-based learning and demonstration strategies of teaching are potent in increasing student achievement. Through project-based learning, students can be motivated to explore, negotiate, interpret, and create in collecting data using the scientific method and formulating the concepts (Olatoye & Adekoya, 2010).There are many factors which influence student achievement. Those factors, according to Beyessa (2014), include parental involvement, peer pressure, schools support and other stakeholders' commitment to improving students' science education. Moreover, the lack of laboratory chemicals, rooms, apparatuses, technicians and well-organized laboratory manuals negatively affect the effective implementation of science education and students' academic achievement.

The implementation of the scientificmethod in teaching science may become a research method. For example, Çaparlar & Dönmez (2016) implemented scientific research forthe purpose of contribution towards sciencethrough systematic collection, interpretation,and evaluation of data in a planned manner. Therefore, students should be taught scientific methods so that they can master research methods.

One of the main issues in classroom assessment is the application for diagnostic and formative purposes. In diagnosed student learning, misconceptions are one of the big issues. Misconceptions may occur because students try to comprehend their previous experiences based on their interactions with their environment (Cañada et al., 2017). Thus, misconceptions on scientific knowledge can be identified if the measurement is not associated with the learning process which has just taken place. In this case, for the Natural Science subject at elementary school, the tests must measure student achievement onthe scientific product (scientific knowledge) and process skills (scientific methods aspects). If this is implemented, the result of the assessment can be used to improve learning. This is relevant with Black et al. (2004) who suggest that an assessment for learning should be interpreted as the teachers' effortsin improving students' learning by utilizing different assessment results, i.e. assessments which are carried out from time to time, from meeting to meeting, and from day to day. Therefore, it is called a mechanism of assessment for learning because the results of measurements on the achievement of scientific knowledge are used as an input for learning.

The cognitive aspects developed in the 2013 Curriculum refer to the principles of Bloom's taxonomy as proposed by Anderson & Krathwohl (2001). These principles explain that scientific knowledge consists of facts, concepts, and procedures (factual, conceptual, and procedural knowledge). In science, the achievement of procedural knowledge is essential because it enables students to practice scientific procedures.

The scientific method is the process by which scientists conduct investigations to produce scientific products (Carin & Sund, 1989). LeBoffe & Wisehart (1989) state that the scientific method is a science process skill which is systematically arranged to solve a problem. In regards to the scientific process and scientific attitude, scientists discover scientific products such as facts, concepts, and new principles in science (Carin & Sund, 1989). Scientific procedures consist of stages of (a) identifying problems; (b) gathering facts and information; (c) designing investigations; (d) reporting the results; (e) repeating investigations/ experiments; (f) checking the results; and (g) drawing conclusions (Belardo & Samia, 1999). The science teaching process to 15 years-old students must encourage them to design and carry out investigations, create data, and interpret the data (Watts et al., 1989). Therefore, teaching scientific procedures to the students must be performed in stages.

Bryce et al. (1995) state that scientific methods consist of a number of basic and process skills that will produce investigative skills when they are arranged into systematic stages. According to Rezba et al. (2007), science process skills are basic skills. When they are integrated,

they will generate integrative skills in the form of scientific methods. According to Bryce et al. (1995), the aspects of basic skills consist of some sub-aspects skills,namely (a) observing using senses; (b) recording data/ information; (c) following instructions; (d) classifying; (e) measuring; (f) manipulating movements; (g) implementing procedures and using equipment; and (h) predicting. Meanwhile, process skills include sub-aspect skills of (a) inferencing and (b) selecting procedures. However, Subali (2009) has indicated that the sub-aspect of "predicting" should be classified in the aspect of process skills. Moreover, Subali (2009) adds that skills of conducting investigation include the sub-aspect skills of (a) planning investigation; (b) conducting investigations; and (c) reporting the findings either in the written or spoken form.

According to Wenning (2010), there are six levelsof inquiry: (a) discovery learning (developing concepts on the basis of first-hand experiences, introducing terms); (b) interactive demonstration (eliciting, identifying, confronting, and resolving alternative conceptions); (c) inquiry lesson (identify scientific principles and/or relationships); (d) inquiry labs (establishing empirical laws based on measurement of variables; (e) real-world applications (applying prior knowledge to authentic problems); and (f) hypothetical inquiry (deriving explanations for observed phenomena). The skills were developed by discovery learning as the lowest level inquiry.They belong to rudimentary skills which includeobserving, formulatingconcepts, estimating, drawing conclusions, communicating results, and classifying results.

Arlianty et al. (2017) state that the scientific method was firstly introduced in the United States in educational science in the 19th century by emphasizing the laboratory formalistic methods which direct scientific facts. The scientific method has the characteristic of "hands-on science." This method enables teachers and curriculum developers to improve the learning process. According to McInerney (1986), The Biological Science Curriculum Study was established in 1958 as a non-profit corporation which is responsible to restructure the Biology Curriculum. The results were initially introduced in 1970 entitled "The Antiquated Content of the Biology Courses in High School" with more experimental, quantitative content. Moreover, the designed materials promoted the teaching of science as a process of investigation and inquiry, rather than in a simple and holistic manner as a body of knowledge.

In Indonesia, the scientific method in the teaching of science should be introduced to learners as early as possible.This has been stated in the 2013 Curriculum for Natural Science subject inthe elementary school level.Since it consists of a series of complex scientific process, its teaching must be carried out little by little, aspect by aspect, and even sub-aspect by sub-aspect. Thus, teaching the scientific method to elementary school students means teaching the students to acquire every aspect and sub-aspect of science process skills.

The tests for measuring the student achievement based on the scientific method teaching may take form as a performance test in the form of work sample tests and paper-pencil/written test. A paper-pencil test can measure the achievement of students' cognitive skills in the scientific method. This is one aspect that must be measured to examine student achievement based on the use of scientific method in teaching science. This study intends to investigate the status of elementary school student achievement based on scientific method, in relation to the test item types as reflected by item difficulty indexes based on the classical test theory (CTT) and modern test theory (IRT). The first stage in developing test was preparing the learning continuum of scientific method aspects by referring to the learning continuum.

A good test development ideally refers to a learning continuum (Northwest Evaluation Association, 2001) so that it can measure the students' skills. The learning continuum is also called a learning trajectory. It refers to student ability which ranges from the lowest to the highest ability related to the complexity and difficulty of content knowledge. A number of studies have discussed the topic of how to develop a written test,such as those by Millard (2012), Miller (2008), Popham (2005), Gronlund & Linn (1990), Gronlund (1998), as well as Roid & Haladyna (1982). Meanwhile, research on test types has been performed by Luo and Zhang (2011). In addition, Shete et al. (2015) conducted research on item analysis for evaluating multiple choice questions in a physiology examination.

In constructing a test, one must consider the use of language. Sumantri & Satriani (2016) find that most teachers agree that communication in the process of teaching and learning is very important to improve students' understanding. Effective communication increases students' understanding of the topics being taught, which

enables them to overcome high-level problems correctly. Effective language is important in the construction of test items so that students can understand the questions easily. Non-standard test items are more difficult for students to answer correctly than the standard test items, as it provides no enhanced ability to discriminate between higher and lower-performing students, resulting in poorer student performance. With regard to this, item-writing guidelines should be taken into account during test construction (Caldwell & Pate, 2013).

In some studies compiled by Ulu (2017), the procedural knowledge may be insufficient in first-time situations, and contextual knowledge is needed for such situations. With regard to this, when the measurement of the achievement of the scientific method is taken, its success may rely on the testee's experience in applying the scientific method and context encountered when they utilize a scientific method.

The quality of the test must meet the theoretical validity viewed from the aspects of construct validity, content validity, and empirical validity. In this case, validity can be viewed from classical test theory (CTT) and modern test theory/Item Response Theory (IRT). The application of modern test theory for analyzing test items has been carried out by Le (2013).

According to Le (2013) whoreferred to Osteen, IRT is considered as the standard validity test. A lot of testing programs still refers to CTT in their design and assessment of test results. This is due to some advantages of CTT over IRT. For example, CTT explains the relationship between the true score and observed score in a linear fashion which makes the CTT model easy to understand and is applicable for a lot of researchers. CTT also offers smaller sample sizes which are smaller than IRT. Compared to IRT, CTT's mathematical procedures are much simpler. In CTT, parameter estimation is conceptually straightforward and requires minimum assumptions,making the model useful and widely applicable. CTT analyses do not need strict goodness of fit studies like that of IRT. However, CTT has somemain weaknesses. One of them is that the test scores rely on the testees. It means that the examinees can get a better score on easier tests and bad scores on difficult tests. Thus, there is no real score which can be extracted because there is no information about the examinees' abilities. This does not allow the test items to matchwith ability levels.

By using IRT, the children's ability level and item difficulty index can be plotted in a single line using a logit scale. Thus, the difficulty level of the items can be compared to the ability of the testee. Meanwhile, CTT item difficulty level cannot be compared to the student's ability (Wright, 1999; Wright & Masters, 1982).

Research on the empirical characteristics of items that attempts to compare CTT and IRT approaches has been widely performed. For instance, Stage (2003) conducted a research on the experience of using CTT and IRT among Swedish examinees. Petrillo et al. (2015) carried out research on utilizing the measurementtheories of CTT, IRT, and Rasch to evaluate the results of patience measurement. Thorpe & Favia (2012) conducted research on the data analysis using IRT methodology for selected programs and applications. Zoghi & Valipour (2014) conducted a study on the comparison of CTT and IRT to predict an item test parameter in linguistics tests. The research investigating the characteristics of items viewed from the characteristics of classical and modern tests theory in relation to scientific method or science process skills have been performed by Pada et al. (2016).

The achievement scientific method can be measured by performance testas well as by paper-pencil or written performance test. The problem in developing an achievement test on scientific methods is the extent to which the written performance test of scientific achievement methods influences the characteristics of the test type based on CTT and IRT. In this case, the problems are (a) whether the multiple choice test with two options has a different difficulty index from those with three options based on CTT and IRT; (b) whether the difficulty index of the analysis performed using CTT is the same as that using IRT; and c) whether the items of a true-false test type have different characteristicsbased on CTT and IRT when the answers are opposite (if a statement is categorized as true in a true-false test of A model, the statement in B model is changed into the false category). Therefore, the aim of this study was to explore studentachievement based on the use of scientific method in teaching science, as reflected on the item difficulty indices and the test item types, without investigating the learning performance in the classroom. In that regard, this study attempts to analyze the problems in the use of scientific method in the teaching and learning

process using to improve its quality in future teaching and learning process.

## METHODS

The test was designed by developing the learning continuum of the scientific aspects method based on the learning continuum of science process skill developed by Subali & Mariyam (2013) and validated by eight (8) experts on Natural Sciences Education from Universitas Sebelas Maret Surakarta and Universitas Negeri Yogyakarta.

The population of this study was elementary school students in Yogyakarta and Sleman Regency in the 2016-2017 school year. The sample consisted of two Technical Management Units (TMUs) in Yogyakarta and one TMU in the Sleman Regency, i.e. TMU of Kalasan sub-district. Three TMUs werechosen by observing the distribution of school locations from the center to the suburbs. They were northern Yogyakarta TMU in the center, eastern Yogyakarta TMUextendingto the suburb, and Kalasan district TMU which was partly located in the countryside. Among those three TMUs, 13 elementary schools were taken from each as samples which were categorized intolow, medium, and high level of performance based on the assessment of the supervisors. Then, one class from Grade 1 to Grade 6 was taken from each school. If a school had a parallel class, all the parallel classes in the school were taken as samples. Thus, each TMU consisted of 78 classes. Therefore, the total sample was 234 classes, in which the number of students for each class ranged from 20 to 30 students. This sample was expected to represent the hypothetical population of the students which may be relevant to another province with the same characteristics. This number of sample was also expected to meet the requirement for testing, which was analyzed using the Graded Model in which the minimum number of examinee required for research is 250 examinees (Muraki & Bock, 1998).

To get a description of the student achievement of Grade 1-6 in elementary school, each test set type was tested to all students.Therefore, in every class, each student was tested using one model of the scientific method achievement test in the form of a paper-and-pencil performance test according to Gronlund'sterminology (1998). Due to the many aspects and sub-aspects of the scientific method, the tests were limited to basic and process skills.

Students in each class were to answer 8 test sets that had been previously prepared. Students who sat side by side were to answer questions with different codes. With regard to this, the results were accountable for the Rasch model. In order to compare the results among students of Grade 1-6, two pairs of the test were provided with an item anchor, for example, code test 1-3 for two options and code test 2-4 for three options. Therefore, a score would be gained within a single measurement scale. The number of the anchorwas 15% of the total items. The item analysis was performed using the Quest Program (Adams & Kho, 1996) to get the results based on the CTT and IRT models.

Since the test was administered simultaneously, different tests were developed in order to prevent the examinees who sat side by side to have the same test. The tests were developed in two different models namely a true-false and multiple choice. The multiple choice test type consisted of test items with two options and those with three options. For the true-false test, there were two models of which the key answers are opposite. If a statement of the true-false test model coded A was declared true, the same statement was changed into a false statement on the test coded B. The test item development considers the aspects of substance, development, and language.

Relevant to the characteristics of natural sciences, each sub-aspect/ indicator of the scientific method can be related to both living and non-living objects. Each test set consisted of 35 items in which 20 items were related to living objects and 15 items were interrelated to non-living objects. The sub-aspects of the scientific method were divided into four sub-aspects, namely sub-aspects I, II, III, and IV which covered basic and process skills. Some test sets contained sub-aspects I and III, and others contained sub-aspects of II and IV. Groups with sub-aspects I and II were related to living objects, and those with sub-aspects III and IV were related to non-living objects. They are presented in Table 1. Moreover, Table 1 presents the result of test analysis using a quest program to get information about test validity and reliability. Test validity is achieved using IRT information in which the test is declared to be "valid" based on classical theory if each item fit the Parameter Logistic (PL) model in IRT. In other words, the test is "valid" based on classical test theory (Wright & Master 1982). In addition, the table presents the amount of reliability index which is expressed in the form of error of measurement error and internal consistency.

**Table 1.** Types of the Test Set on Student Achievement Based on Scientific Method Teaching along with the Specification of Questions, Sub-Aspects, Scientific Methods, and Natural Objects

| Types of the test set | Types of question | Sub-Aspect Groups of Scientific Method | Specification of Items Related to the Objects | Items Fit With the Rasch Model | Reliability (Error of Measurement/Internal Consistency) |
|---|---|---|---|---|---|
| Test set 1 | Multiple choice type with two answer choices | Group of Code I<br>Group of Code III | 20 items are related to living thing objects (item number 1 to 20)<br>15 items are related to non-living thing objects (item number 21 to 35) | All items fit | 0.72/0.72 |
| Test set 2 | Multiple choice type with three answer choices | Group of Code I<br>Group of Code III | 20 items are related to living thing objects (item number 1 to 20)<br>15 items are related to non-living thing objects (item number 21 to 35) | All items fit | 0.70/0.70 |
| Test set 3 | A true-false type of A model | Group of Code I<br>Group of Code III | 20 items are related to living thing objects (item number 1 to 20)<br>15 items are related to non-living thing objects (item number 21 to 35) | All items fit | 0.51/0.49 |
| Test set 4 | A true-false type of B model (the key answer is opposite to that of A model) | Group of Code I<br>Group of Code III | 20 items are related to living thing objects (item number 1 to 20)<br>15 items are related to non-living thing objects (item number 21 to 35) | All items fit | 0.59/0.57 |
| Test set 5 | Multiple choice type with two answer choices | Group of Code II<br>Group of Code IV | 20 items are related to living thing objects (item number 1 to 20)<br>15 items are related to non-living thing objects (item number 21 to 35) | All items fit | 0.70/0.72 |
| Test set 6 | Multiple choice type with three answer choices | Group of Code II<br>Group of Code IV | 20 items are related to living thing objects (item number 1 to 20)<br>15 items are related to non-living thing objects (item number 21 to 35) | All items-fit | 0.71/0.71 |
| Test set 7 | A true-false type of A model | Group of Code II<br>Group of Code IV | 20 items are related to living thing objects (item number 1 to 20)<br>15 items are related to non-living thing objects (item number 21 to 35) | All items fit | 0.54/0.51 |
| Test set 8 | A true-false type of B model (the key answer is opposite to that of A model) | Group of Code II<br>Group of Code IV | 20 items are related to living thing objects (item number 1 to 20)<br>15 items are related to non-living thing objects (item number 21 to 35) | All items fit | 0.55/0.53 |

**Note:** Based on a quest program, an item fitsthe model if it is in between the InfitMNSQ range of 0.7-1.3.

## RESULTS AND DISCUSSION

The following is the result of data analysis using the classical theory to determine the item difficulty index on the aspects and sub-aspects of scientific methods based on the testing results on a group of examinees of Grade 1 to 3 and Grade 4 to 6.

**Table 2.** The Test Difficulty Index Related to the Aspects and Sub-Aspects of Scientific Method Using CTT Analysis for Sub-Aspects with Indicators I and III

| Code | Indicators of Scientific Processes | Grade of Students | DI of MC | DI of MC | DI of TF | DI of TF |
|---|---|---|---|---|---|---|
| | | | Two Options | Three Options | A Model | B Model |
| 1.1 | Observing skills (4 items) | 1 – 3 | 0.7 | 0.57 | 0.61 | 0.64 |
| | | 4 – 6 | 0.8 | 0.66 | 0.54 | 0.73 |
| 1.2 | Recording data skills using sense of sight/ hearing/smell/taste/touch), (4 items) | 1 – 3 | 0.53 | 0.4 | 0.71 | 0.42 |
| | | 4 – 6 | 0.72 | 0.59 | 0.71 | 0.42 |
| 1.3 | Following instruction skills (1 item) | 1 – 3 | 0.37 | 0.19 | 0.7 | 0.48 |
| | | 4 – 6 | 0.45 | 0.17 | 0.7 | 0.48 |
| 1.4 | Classifying skills (2 items) | 1 – 3 | 0.38 | 0.63 | 0.6 | 0.84 |
| | | 4 – 6 | 0.65 | 0.58 | 0.6 | 0.84 |
| 1.5 | Measuring skills (6 items) | 1 – 3 | 0.49 | 0.36 | 0.62 | 0.45 |
| | | 4 – 6 | 0.61 | 0.47 | 0.62 | 0.45 |
| 1.6 | Manipulating movement skills (5 items) | 1 – 3 | 0.61 | 0.4 | 0.43 | 0.76 |
| | | 4 – 6 | 0.74 | 0.54 | 0.43 | 0.76 |
| 1.7 | Implementing procedures/techniques/ equipment usage skills (4 items) | 1 – 3 | 0.57 | 0.33 | 0.47 | 0.57 |
| | | 4 – 6 | 0.65 | 0.44 | 0.47 | 0.57 |
| 2.1 | Inferencing skills (5 items) | 1 – 3 | 0.65 | 0.61 | 0.81 | 0.6 |
| | | 4 – 6 | 0.83 | 0.82 | 0.82 | 0.61 |
| 2.2 | Predicting skills (3 items) | 1 – 3 | 0.73 | 0.5 | 0.62 | 0.89 |
| | | 4 – 6 | 0.88 | 0.58 | 0.62 | 0.89 |
| 2.3 | Selecting procedure skills (2 items) | 1 – 3 | 0.58 | 0.67 | 0.63 | 0.9 |
| | | 4 – 6 | 0.66 | 0.91 | 0.63 | 0.9 |

Note: DI: Difficulty Index; MC: Multiple choice; TF: True-false

Based on the CTT theory, the lowest the difficulty index is, the easier the item will be. Normally, a certain skill is easier for the students of Grade 4 to 6 than for those of Grade 1 to 3. This is because Grade 4 to 6 students have more learning experience and apply more scientific method aspects than Grade 1 to 3 students. Based on the results of analysis using CTT for indicators I-III, observing skill is more difficult for the Grade 4 to 6 students than for Grade 1 to 3 students when being tested using a true-false test type A. In addition, recording data/information and classifying skills are more difficult for the Grade 4 to 6 students than for the Grade 1 to 3 students when being tested using a multiple-choice test with three options.

**Table 3.** The Test Difficulty Index Related to The Aspects and Sub-Aspects of Scientific Method Using a CTT Analysis for Sub-Aspects with Indicators II and IV

| Code | Indicators | Grade of Students | DI of MC | DI of MC | DI of TF | DI of TF |
|---|---|---|---|---|---|---|
| | | | Two Options | Three Options | A Model | B Model |
| 1.1 | Observing skills (4 items) | 1 – 3 | 0.65 | 0.65 | 0.58 | 0.5 |
| | | 4 – 6 | 0.76 | 0.83 | 0.72 | 0.57 |
| 1.2 | Recording data/information skills (5 items) | 1 – 3 | 0.48 | 0.4 | 0.5 | 0.53 |
| | | 4 – 6 | 0.58 | 0.42 | 0.56 | 0.53 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 1.3 | Following instruction skills (2 items) | 1 – 3 | 0.39 | 0.38 | 0.66 | 0.37 |
| | | 4 – 6 | 0.45 | 0.63 | 0.64 | 0.48 |
| 1.4 | Classifying skills (2 items) | 1 – 3 | 0.47 | 0.37 | 0.73 | 0.35 |
| | | 4 – 6 | 0.72 | 0.49 | 0.8 | 0.51 |
| 1.5 | Measuring skills (4 items) | 1 – 3 | 0.69 | 0.44 | 0.55 | 0.52 |
| | | 4 – 6 | 0.81 | 0.59 | 0.69 | 0.64 |
| 1.6 | Manipulating movement skills (2 items) | 1 – 3 | 0.73 | 0.35 | 0.66 | 0.52 |
| | | 4 – 6 | 0.89 | 0.43 | 0.64 | 0.65 |
| 1.7 | Implementing procedures/techniques/ equipment usage skills (7 items) | 1 – 3 | 0.55 | 0.43 | 0.56 | 0.48 |
| | | 4 – 6 | 0.71 | 0.56 | 0.6 | 0.56 |
| 2.1 | Inferencing skills (5 items) | 1 – 3 | 0.64 | 0.52 | 0.48 | 0.74 |
| | | 4 – 6 | 0.8 | 0.63 | 0.61 | 0.8 |
| 2.2 | Predicting skills (1 item) | 1 – 3 | 0.56 | 0.6 | 0.87 | 0.86 |
| | | 4 – 6 | 0.57 | 0.83 | 0.95 | 0.93 |
| 2.3 | Selecting procedure skills (3 items) | 1 – 3 | 0.56 | 0.49 | 0.43 | 0.49 |
| | | 4 – 6 | 0.69 | 0.59 | 0.49 | 0.57 |

Note: DI: Difficulty Index; MC: Multiple choice; TF: True-false

Based on the CTT theory, the higher the difficulty index is, the easier the item will be.

The results of the analysis using the CTT approach for the indicators II-IV presented in Table 3 show that there is only one aspect i.e. manipulating movement skill, which shows an opposite condition in which it is more difficult for the students of Grade 4 to 6 than for those of Grade 1 to 3 when being tested using a true-false test type A.

**Table 4.** The Test Difficulty Index Related to the Aspects and Sub-Aspects of Scientific Method Using An IRT Analysis for Sub-Aspects with Indicators I and III

| Code | Indicators | Grade of Students | DI of MC | DI of MC | DI of TF | DI of TF |
|---|---|---|---|---|---|---|
| | | | Two Options | Three Options | A Model | B Model |
| 1.1 | Observing skills (4 items) | 1 – 3 | -0.78 | -0.7 | 0.19 | -0.49 |
| | | 4 – 6 | -0.8 | -0.76 | 0.19 | -0.49 |
| 1.2 | Recording data/information skills (4 items) | 1 – 3 | 0.33 | 0.33 | -0.78 | 1.16 |
| | | 4 – 6 | 0.22 | 0.12 | -0.78 | 1.16 |
| 1.3 | Following instruction skills (1 item) | 1 – 3 | 1.01 | 1.37 | -0.3 | 0.86 |
| | | 4 – 6 | 1.51 | 2.22 | -0.3 | 0.86 |
| 1.4 | Classifying skills (1 item) | 1 – 3 | 0.98 | -0.74 | 0.17 | -0.93 |
| | | 4 – 6 | 0.6 | 0.21 | 0.17 | -0.93 |
| 1.5 | Measuring skills (6 items) | 1 – 3 | 0.49 | 0.52 | 0.13 | 0.97 |
| | | 4 – 6 | 0.7 | 0.72 | 0.05 | 0.97 |
| 1.6 | Manipulating movement skills (5 items) | 1 – 3 | -0.14 | 0.28 | 0.94 | -0.68 |
| | | 4 – 6 | -0.24 | 0.36 | 0.94 | -0.68 |
| 1.7 | Implementing procedures/techniques/ equipment usage skills (4 items) | 1 – 3 | 0.13 | 0.59 | 0.72 | 0.49 |
| | | 4 – 6 | 0.52 | 0.84 | 0.72 | 0.49 |
| 2.1 | Inferencing skills (5 items) | 1 – 3 | -0.2 | -0.71 | -0.96 | 0.28 |
| | | 4 – 6 | -0.46 | -1.206 | -0.96 | 0.28 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 2.2 | Predicting skills (3 items) | 1 – 3 | -0.69 | -0.16 | -0.24 | -1.74 |
| | | 4 – 6 | -1.16 | 0.2 | -0.24 | -1.74 |
| 2.3 | Selecting procedure skills (2 items) | 1 – 3 | 0.095 | -0.99 | 0.07 | -1.57 |
| | | 4 – 6 | 0.45 | -1.95 | 0.07 | -1.57 |

Note: DI: Difficulty Index; MC: Multiple choice; TF: True-false

Based on the IRT theory, the higher the difficulty index is, the more difficult the item will be.

The results of the analysis using IRT for indicators I-III as presented in Table 4 show that some indicators are more difficult for the students of Grade 4 to 6 than for those of Grade 1 to 3 when being tested with multiple-choice tests of two options and multiple-choice tests of three options. Those indicators for multiple-choice tests with two options include following instructions, measuring, and implementing procedures/techniques/equipment usage skills. Moreover, for multiple-choice tests with three options, the indicators include those 3 aforementioned skills and the skills of classifying and manipulating movement.

**Table 5.** The Test Difficulty Index Related to the Aspects and Sub-Aspects of Scientific Method Using an Analysis of Modern Test Theory for Sub-Aspects with Indicators II And IV

| Code | Indicators | Grade of students | DI of MC two options | DI of MC three options | DI of TF A model | DI of TF B model |
|---|---|---|---|---|---|---|
| 1.1 | Observing skills (4 items) | 1 – 3 | -0.34 | -0.93 | -0.1 | 0.18 |
| | | 4 – 6 | -0.21 | -1.49 | -0.34 | 0.26 |
| 1.2 | Recording data/information skills (5 items) | 1 – 3 | 0.43 | 0.29 | 0.27 | 0.04 |
| | | 4 – 6 | 0.74 | 0.89 | 0.4 | 0.43 |
| 1.3 | Following instruction skills (2 items) | 1 – 3 | 0.86 | 0.42 | -0.43 | 0.72 |
| | | 4 – 6 | 1.39 | -0.1 | 0.05 | 0.67 |
| 1.4 | Classifying skills (2 items) | 1 – 3 | 0.55 | 0.53 | -0.74 | 0.85 |
| | | 4 – 6 | 0.09 | 0.59 | -0.74 | 0.57 |
| 1.5 | Measuring skills (4 items) | 1 – 3 | -0.55 | 0.14 | 0.07 | 0.07 |
| | | 4 – 6 | -0.58 | -0.04 | -0.16 | -0.11 |
| 1.6 | Manipulating movement skills (2 items) | 1 – 3 | -0.69 | 0.56 | -0.45 | 0.12 |
| | | 4 – 6 | -1 | 0.89 | 0.02 | -0.1 |
| 1.7 | Implementing procedures/techniques/ equipment usage skills (7 items) | 1 – 3 | 0.16 | 0.14 | 0.05 | 0.25 |
| | | 4 – 6 | 0.11 | 0.25 | 0.19 | 0.29 |
| 2.1 | Inferencing skills (5 items) | 1 – 3 | -0.3 | -0.28 | 0.32 | -1 |
| | | 4 – 6 | -0.6 | -0.3 | 0.04 | -1.09 |
| 2.2 | Predicting skills (1 item) | 1 – 3 | 0.11 | -0.57 | -1.66 | -1.7 |
| | | 4 – 6 | 0.87 | -1.2 | -2.34 | -2.1 |
| 2.3 | Selecting procedure skills (3 items) | 1 – 3 | 0.08 | -0.12 | 0.58 | 0.15 |
| | | 4 – 6 | -0.05 | -0.06 | 0.68 | 0.18 |

Note: DI: Difficulty Index; MC: Multiple-choice; TF: True-false

Based on the IRT theory, the lowest the difficulty index is, the more difficult the item will be. The results of the analysis using IRT for indicators II-IV presented in Table 5 show that some indicators are more difficult for the Grade 4 to 6 students than for the Grade 1 to 3 students

when being tested with a multiple-choice test of two options, multiple-choice test of three options, true-false test type A, and true-false test type B. Those indicators include the indicators for multiple-choice tests with two options such as observing, recording data/information, following instructions, and predicting skills; the indicators for multiple-choice tests with three options such as recording data/information, classifying, and selecting procedure skills; the indicators for the true-false test type A such as recording data/information, manipulating movement, implementing procedures/techniques/equipment usage, and selecting procedure skills; and the indicators for the true-false test type B such as observing, recording data/information, implementing procedures/techniques/equipment usage, and selecting procedure skills.

The results show that indicators I-III and II-IV signify a different index of difficulty regarding the student achievement in the scientific method aspects both in basic and process skills when being tested with a different test type. Although the indicator is the same, the difficulty index will be different when being tested with a different test type. In addition, a multiple-choice test with threeoptions is not always more difficult than that with two options. Similarly, a multiple-choice test with threeoptions is not always more difficult than a true-false test type. This implies that designing a test by changing the test types is not reliable. However, further research is needed. Although an IRT is employed, it will not be affected by who the examinee is (Le, 2013).

The second finding is that the difficulty index for Grade 4 to 6 students becomes higher than that of Grade 1 to 3 students. This indicates that learners who have more learning experiences do not always have better achievement in science process skills. This occurs because the skills of the scientific process as parts of the scientific method are not taught optimally to the students. Some of the supervisors involved in community service activities repeatedly state that the main focus of the teaching is the student achievement in the national examination. Thus, the national examination is regarded as a high-stakes test. When a test is viewed as a high-stakes test, the teachers are possibly focusing on the students' success to deal with it and may lead to teaching for the test. Criticisms about the implementation of the test, both national tests and state tests, addressed to the government are also found in developed countries like the United States (US). Both tests are also considered as a high-risk test. This is due to the many failures experienced by students when taking the national and state test in Elementary and Secondary Education Act program or No Child Left Behind (NCLB) in 2001. Criticisms about these tests are described in the articles written by Garrison (2009). Almost all states, schools, and teachers focus on pursuing a learner's success in dealing with the state tests. Many praxes of learning focus on the students' better achievement in the test, or it is often called teaching for the test.

The third finding also indicates that the implementation of the 2013 Curriculum does not give an impact on the student achievement in skills, including their achievement in scientific method aspects in the teaching of Natural Science subjects at elementary schools. This is because the teachers, supervisors, and even the Principal of TMU keep focusing on the student achievement in the National Examination even though this examination does not measure the student achievement in skills which are parts of the scientific method aspects. Therefore, it is necessary to conduct research on how the scientific method is actualized in learning Natural Science subjects in elementary schools.

The importance of conducting research on the actualization of scientific learning performed in an inquiry model is based on the literature of science literacy which encourages teachers to employ inquiry as a regular part of teaching practice (e.g., National Science Education Standards, Science for All Americans: Project 2061). Unfortunately, this does not always happen. In summary, the success of those 6 inquiry levels depends on the teacher's ability (Wenning, 2010). The low capability of teachers in teaching scientific method is due to their low understanding of the Nature of Scientific Inquiry (NOSI). It happens because the teachers do not understand how to do it when they were taking the study. Miller et al. (2010) state that pre-service and in-service training enables teachers to possess informed conceptions about NOSI. With these informed conceptions, teachers may internalize the instructional importance of NOSI which, in turn, may help avoid the lack of attention to NOSI currently evidenced in teachers' instructional decisions. This might result in teachers' orientation shifting towards an explicit inquiry-based approach from that of an implicit science process and discovery approach. In addition, Hairida & Junanto (2018) say that the low science literacy skills of students in Indonesia may be influenced by several factors, such as the instructional model applied by teachers and the teaching materials used by the students. Moreover, Dudu (2014) says that students' scores on sections that address the six aspects of Nature

of Science (NOS) were significantly different in most cases, showing notably uninformed views of the distinctions between scientific theories and laws. Evidence-based insight into students' NOS views can aid in reforming undergraduate science courses and will add to faculty and researcher understanding of the impressions of science held by undergraduates, helping educators improve the scientific literacy of future scientists and diverse college graduates. Based on the discussion above, the test results can be used as an input for teachers to improve the quality of learning which focuses on the learning which introduces the components of the scientific method to train the students to find new facts and concepts.

## CONCLUSION

The conclusion of this research is that different types of tests show a different index of difficulty when measuring student achievement in scientific method aspects, including basic skills and process skills. In addition, there are some aspects of the scientific method, both basic and process skills, that are even more difficult for Grade 4 to 6 students than for Grade 1 to 3 students. The student achievement based on the scientific method reflected by item difficulty index based on CTT and IRT indicates the same level of category, except for few sub-aspects. The data also show that few items of certain tests indicate higher difficulty for Grade 4 to 6 students than for Grade 1 to 3 students for the same tests. This case is not relevant to the expected learning outcomes. Based on this summary, the actualization of scientific method-based learning in Natural Science subjects in elementary schools should be investigated.

## ACKNOWLEDGMENTS

The researchers thank the Directorate of Research and Community Service of the Ministry of Research, Technology, and Higher Education which has supported the researchers by providing financial support so that this study could be carried out.

## REFERENCES

Adams, R.J. & Kho, Seik-Tom. (1996). *Acer Quest Version 2.1*. Camberwell, Victoria: The Australian Council for Educational Research.

Anderson, L.W. &Krathwohl, D.R. (2001). *A Taxonomy of Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives.* A Bright Edition. New York: Addison Wesley Longman, Inc.

Arlianty, W.N, Febriana, B.W, & Diniaty, A. (2017). An Analysis of Learning Process Based on Scientific Approach in Physicalchemistry Experiment. *AIP Conference Proceedings* 1823, 020084 (2017).

Belardo, P.S. & Samia, E.R. (1999). *Integrated Science & Technology I: Laboratory Manual*. Quezon City: FNB Educational, INC.

Beyessa, F. (2014). Major Factors that Affect Grade 10 Students' Academic Achievement in Science Education at Ilu Ababora General Secondary of Oromia Regional State, Ethiopia. *International Letters of Social and Humanistic Sciences, 32*, 118-134.

Black, P., Harrison, Ch., Lee, Cl., Marshall, B., & Dylan, W. D. (2004). *Assessment for Learning: Putting it into practice.* New York: Open University Press.

Bryce, T.G.K., McCall, J., MacGregor, J., Robertson, I.J., & Weston, R.A.J. (1995). *Techniques for assessing process skills in practical science: Teacher's guide*. Oxford: Heinemann Instructional Books.

Caldwell, D.J. & Adam N. Pate, A.N. (2013). Effects of Question Formats on Student and Item Performance. *Am J Pharm Educ*, *77*(4), 71.

Cañada, Fl.C., González-Gómez, D., Airado-Rodríguez, D., Niño, L.V.M., & Acedo, M.A.D. (2017). Change in Elementary School Students' Misconceptions on Material Systems after a Theoretical-Practical Instruction. *International Electronic Journal of Elementary Education*, *9*(3), 499-510.

Çaparlar, C. Ö., & Dönmez, A. (2016). What is Scientific Research and How Can It Be Done? *Turkish Journal of Anesthesiology and Reanimation*, *44*(4), 212.

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

Dudu, W.T. (2014). Exploring South African High School Teachers' Conceptions of Thenature of Scientific Inquiry: A Case Study. *South African Journal of Education, 34*(1),1-19. Retrieved from: http://www.sajournalofeducation.co.za

Garrison, M. J. (2009). *A Measure of Failure: The Political Origins of Standardized Testing*. Albany: SUNY Press.

Gronlund, N.E. (1998). *Assessment of Student Mastery* (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.

Hairida & Junanto, T. (2018). The Effectiveness of Performance Assessment in Project-Based Learning by Utilizing Local Potential to Increase Science Literacy. *International Journal of Pedagogy and Teacher Education*, *2*, 151-162.

Le, Dai-Trang. (2013). *Applying Item Response Theory Modeling in Educational Research*. (Doctoral Dissertations, Iowa State University). Retrieved

from: http://lib.dr.iastate.edu/etd

LeBoffe, M. & Wisehart, G. (1989). *Study Guide Biology: Exploring Life*. New York: John Wiley & Sons.

Lederman, N. G., Lederman, J. S., & Antink, A. (2013). Nature of Science and Scientific Inquiry as Contexts for the Learning of Science and Achievement of Scientific Literacy. *International Journal of Education in Mathematics, Science and Technology*, *1*(3), 138-147.

Luo, S.& Zhang, X. (2011). Multiple-choice Item and Its Backwash Effect on Language Teaching in China. *Theory and Practice in Language Studies, 1*(4), 423-425.

McInerney, J. (1986). Curriculum Development at the Biological Science Curriculum Study. *Educational Leadership: Journal of the Association for Supervision and Curriculum Development*, *44*(4), 24-28.

Millard, S. (2012). *Writing Multiple Choice and True/ False Exam Questions. A Good Practice Guide.* (Lecture Notes). Retrieved from: https://docplayer.net/50079039-Writing-multiple-choice-and-true-false-exam-questions.html

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

Miller, M.C.D., Montplaisir, L.M., Offerdahl, E.G., Cheng, F-Ch., & Gerald L., Ketterling, G.L. (2010). Comparison of Views of the Nature of Science between Natural Science and Nonscience Majors. *CBE—Life Sciences Education*, *9*, 45–54.

Northwest Evaluation Association. (2003). Idaho State Aligned Learning Continuum Release 1.0. *Nweverest StPortland org*, *4*.

Nweke, C.O., Abonyi, O. S., Chinyere A. O., &Njoku, M.I.A. (2014). Effects of Experiential Teaching Method on Pupils' Achievement in Basic Science and Technology. *International Journal of Scientific & Engineering Research*, *5*(5), 875-881.

Olatoye, R.A.and Adekoya, Y.M. (2010). Effect of Project-Based, Demonstration and Lecture Teaching Strategies on Senior Secondary Student achievement in an Aspect of Agricultural Science. *International Journal of Educational Research and Technology*, *1*(1), 19-29.

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), 1-12. Retrieved from: http://journal.uny.ac.id/index.php/reid

Petrillo, J., Cano, S.J., McLeod, L.D., & Coon, Ch.D. (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*, 25 -34.

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*. Orlando: Academic Press, Inc.

Shete, A. N., Kausar, A., Lakhkar, K., & Khan, S. T. (2015). Item Analysis: An Evaluation of Multiple-Choice Questions in Physiology Examination. *J Contemp Med Edu*, *3*(3), 106-109.

Stage, C. (2003). Classical Test Theory or Item Response Theory: The Swedish Experience (PDF file). Retrieved from: https://www.umu.se/globalassets/organisation/fakulteter/samfak/institutionen-for-tillampad-utbildningsvetenskap/hogskoleprovet/publications/60581_emno-42.pdf

Subali, B. (2009). *Pengukuran Keterampilan Proses Sains Pola Divergen dalam Mata Pelajaran Biologi SMA di Provinsi DIY dan Jawa Tengah*. (Unpublished dissertation). Yogyakarta State University, Yogyakarta, Indonesia.

Subali, B., & Mariyam, S. (2013). Pengembangan Kreativitas Keterampilan Proses Sains dalam Aspek Kehidupan Organisme Pada Mata Pelajaran IPA SD. *Cakrawala Pendidikan*, *3*(3),365-381.

Sumantri, M. S., & Satriani, R. (2016). The Effect of Formative Testing and Self-Directed Learning on Mathematics Learning Outcomes. *International Electronic*

Tang, X., Coffey, J. E., Elby, A., & Levin, D. M. (2010). The Scientific Method and Scientific Inquiry: Tensions in Teaching and Learning. *Science Education*, *94*(1), 29-47.

Thorpe, G. & Favia, A. (2012). Data Analysis Using Item Response Theory Methodology: An Introduction to Selected Programs and Applications (PDF file). Retrieved from: https://digitalcommons.library.umaine.edu/cgi/viewcontent.cgi?article=1019&context=psy_facpub

Tsybulsky, D. (2019). Students Meet Authentic Science: The Valence and Foci of Experiences Reported by High-School Biology Students Regarding Their Participation in a Science Outreach Programme. *International Journal of Science Education*, *41*(5), 567-585.

Ulu, M. (2017). Errors Made by Elementary Fourth Grade Students When Modelling Word Problems and the Elimination of Those Errors through Scaffolding. *International Electronic Journal of Elementary Education*, *9*(3), 553-580.

Watts, M., Bentley, D., & Hornsby, J. (1989). *Learning to Make it Your Own. In: Bentley, D. & Watts, M. (ed). Learning & Teaching in Schools Science: Practical Alternative.* Philadelphia: Open University Press.

Wenning, C.J. (2010). Levels of Inquiry: Using Inquiry Spectrum Learning Sequences to Teach Science (Shaded Sections Added January 2012). *J. Phys. Tchr. Educ*, *5*(3), 11-20.

Wright, B.D. (1999). Rasch Measurement Model. In: Masters, G.N. & Keeves, J.P. (1999). *Advances in*

*Measurement in Educational Research and Assessment*. Amsterdam: Pergamon, An imprint of Elsevier Science.

Wright & Masters, G.N. (1982). *Rating Scale Analysis*. Chicago: Mesa Press.

Zeidan, A. H., & Jayosi, M. R. (2015). Science Process Skills and Attitudes toward Science among Palestinian Secondary School Students. *World Journal of Education*, *5*(1), 13-24.

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*(4), 424-435.