LOOK
ACADEMIC PUBLISHERS
OPEN ACCESS

INTERNATIONAL JOURNAL OF ENVIRONMENTAL & SCIENCE EDUCATION
2017, VOL. 12, NO. 8, 1719-1735

Development of Instrument to Assess Cognitive Process and Product in Biology Senior High School

Paidi[a], Djukri[a], Siti Yulaikah[a], Dessy Alfindasari[a]

[a]Yogyakarta State University, Yogyakarta, INDONESIA

ABSTRACT

The aim of this study was to develop the instrument(s) to access the senior high school (SHS) student competence of cognitive process and product dimensions on Biology. The developed instruments were written objective test in form of multiple-choice test and objective description test (multiple-choice test with argumentation). The research modified the instrument item development procedure from L.L. Oriondo & E.M. Dallo-Antonio (2008). The procedure consists of (1) instrument item drafting, and (2) field testing. The item drafting was be done via (a) determining aspects of competence to test, (b) determining a relevant biology subject matter(s), (c) developing table of instrument specification, (d) constructing the instrument items, (e) composing criteria for scoring, (f) reviewing the instrument item(s), and (g) revising. The field testing was being done in April-October 2016 involving 1126 Biology SHS students came from 4 representative provinces of Indonesia as respondent. The field testing data was be analyzed descriptively using QUEST, BILOG, and Parscale to find-out validity and reliability of the instrument items, including the goodness of fit, difficulty index, reliability estimate, item characteristic curve (ICC), Standard Error of Measurement (SEM), and standard error of measurement (SEM). Results of the research showed that the developed instrument was valid and reliable. The objective test items have an INFIT MNSQ 1,00 and standard deviation of 0,04; The objective description test items have an INFIT MNSQ 1,00 and standard deviation of 0,13. Difficulty indexes for the test items were in good criteria (ranged -0.75 to +0,7). Reliabilities of item estimate were 0,94 (for objective test item) and 0.80 (for objective description test item). ICC for almost all objective test items showed the high ability and showed the low ability for almost all objective description test items. SEMs for the item were $-1,5 < \theta < 2,5$ (for objective test) and $-2,7 < \theta < 1,9$ (for objective description test); Its means the items were fit for student with low and high ability.

KEYWORDS
objective and objective description test item, cognitive process and product dimensions, biology SHS

ARTICLE HISTORY
Received 15 February 2017
Revised 23 April 2017
Accepted 19 May 2017

Introduction

The 21st century as a global era is marked by the development of science and technology. Development in science and technology is actually a momentum

CORRESPONDENCE Paidi ✉ paidi@uny.ac.id

to further improve the quality of human resources, including those in Indonesia. Based on a study by the United Nations Development Program in 2013, the Human Development Index of Indonesia ranks 108 out of 187 countries. This indicates that the expectations in the live index, education index, and gender balance index are still undervalued. The results show that the Indonesian human resources need strengthening to face the global era, which is understood as an era of competition.

One international study discusses the students' cognitive abilities are TIMSS (Trends In Mathematics and Science Study) that is held by the IEA (International Associations for the Evaluations of Education Achievement) (2011) showed that Indonesia achieved a value of 5% in the category of work on the problems of reasoning (HOTS), TIMSS implement aspects of the understanding, application and reasoning in cognitive dimension is divided into two ability to think low and high thinking ability. Interest aspect of the learning process that teachers and students can produce learning outcomes that can sharpen the ability to think the issue through several development competencies so that the learning process that teachers need to be an emphasis on student-dimensional thinking, or better known as cognitive dimensions.

One of the competencies needed in the global era is associated is cognitive skills or critical thinking. Yu-Mei-Lin and Pei-Chen Lee (2013) described critical thinking as a skill that requires high levels of skill, knowledge and reconstructs questions in the troubleshooting process. Cognitive Domain is committed to treatment by pulling back and develop intellectual abilities yourself. The cognitive domains are specified into three levels are knowledge, understanding, applying to the higher process. And then Based on the above explanation Ministry of Education, Province of British Columbia (2008) through the Resource Integrated Package describes that the cognitive domain was the recall of knowledge and the development of one's intellectual abilities. Cognitive domains can be said to further include a cognitive three levels, namely: knowledge, understanding, use, and process capability so the highest brain can be used to find out the students' learning achievement.

Theory of cognitive created by B.S. Bloom (1956) is still used as a broad reference in the practice of Indonesian education. In its development, Bloom's taxonomy is revised by L.R. Anderson and D.R. Krathwohl (2001) in which the cognitive domain is divided into two dimensions, i.e. cognitive processes and cognitive products. The cognitive process dimensions are associated with the processes of the six levels or categories (C1 to C6) expressed in the verbs remembering, understanding, applying, analyzing, evaluating, and creating. Knowledge of cognitive aspects the cognitive process dimension hearts are divided into prayer That ability to think low skills (LOTS) consisting of C1 to C3 and higher order thinking skills consisting of C4 to C6. Meanwhile, the cognitive product dimensions consist of four levels or categories, namely factual knowledge, conceptual knowledge, procedural knowledge, and metacognitive knowledge.

According to L.R. Anderson and D.R. Krathwohl (2001), *remembering* consists of recognizing and recalling relevant information from long-term memory. *Understanding* is the ability to make your own meaning from an educational material such as teacher explanations. Applying refers to using a learned procedure either in a familiar or new situation. Analyzing which consists of breaking knowledge down into its parts and thinking about how the parts relate to its overall structure. E*valuating* includes checking and critiquing. *Creating* is the highest and new component of the new version of the cognitive taxonomy. Creating involves putting things together to make something new. To accomplish creating tasks, learners generate, plan, and produce. Factual knowledge includes isolated bits of information, such as vocabulary definitions and knowledge about specific details. Conceptual knowledge consists of systems of information, such as classifications and categories. Procedural knowledge includes algorithms, heuristics or rules of thumb, techniques, and methods as well as knowledge about when to use these procedures.

Metacognitive knowledge refers to knowledge of thinking processes and information about how to manipulate these processes effectively. Meanwhile, according to A. Widodo (2006) explains that knowledge of metacognitive someone will continue to increase with the development of students, thus bringing the students to know the awareness itself to better learning.

According to X.N. Wu, H. Wu, W. Wang (2016) that affect cognitive abilities students against process capability, thus allowing their cognitive abilities in relations to Student Learning. Dettmer (2006) explains that the most important reason for their taxonomic help educators knows about the recall, understand the knowledge and understanding of the teaching and assessment are mastered.

In the new version of Bloom's taxonomy, L.R. Anderson and D.R. Krathwohl (2001) also gave an intersection between and cognitive product (knowledge) and process dimensions (Table 1). The intersection can facilitate teachers and educators in selecting of teaching activities. O.F. Tutkun et al. (2012) agreed to the idea, the intersection enables teachers and educators to identify which knowledge they expect students to use and to determine which cognitive process dimension is used. A learner can remember factual or procedural knowledge, understand conceptual or metacognitive knowledge, or analyze metacognitive or factual knowledge. According to L.R. Anderson and D.R. Krathwohl (2001), "Meaningful learning provides learners with the knowledge and cognitive processes they need for successful problem solving".

Table 1. Placement of The Objective and Instructional Activities in The Taxonomy Table

The Cognitive Product (Knowledge) Dimension	The Cognitive Process Dimension					
	Remember (C1)	Understand (C2)	Apply (C3)	Analyze (C4)	Evaluate (C5)	Create (C6)
Factual Knowledge (K1)						
Conceptual Knowledge (K2)		Activity 1		Activity 2		
Procedural Knowledge (K3)			Activity 3			
Metacognitive Knowledge (K4)	Activity 4					

According to the Table 1 above, there will be 24 intersections between cognitive process dimension (c) and knowledge (cognitive product) dimension (K), namely: C1K1, C1K2, ...C6P4 to become activities of teaching, learning, and assessing.

That is, the use of taxonomy can provide one advantage, namely, there are several behaviors that should be emphasized in the planning education, as also confirmed by A.J. Nitko & S.M. Brookhart (2011). Use of the taxonomy can also help a student gain a perspective on the emphasis given to certain behaviors by a particular set of educational plans. Educators should find the taxonomy helps them to specify objectives so that it becomes easier to plan learning experiences and prepare an evaluation device (Dettmer, 2006).

The urgency of developing the cognitive process and products abilities in senior high school (SHS) students are mentioned in National Standard of Education in Indonesia. Since 2013, Indonesian Government has decided some standards in education. Two of the standards are Standard of content and standard of competency Graduate. According to the standards, the cognitive process and product abilities referring to the Revised Bloom's Taxonomy, must be mastered by SHS student (even by junior high school and elementary school students) through out all subjects.

Biology, one of the subjects studied in SHS, has a role in developing students' ability to think through the learning processes. Suciati (2015) states that the characteristics of the biological material are different from other disciplines, meaning that in biology many things that could be explored include studying the biology of living beings, the environment, and the relationship between them. So it requires a high level of ability in studying every aspect and existing studies in the biological sciences.

As already noted, the massive effort has been made to improve the mastery of cognitive process and product dimensions of learners through a variety of socialization and practice imposed by the government. However, in the processes, no analysis was not carried out concerning the students' levels of achievement on the cognitive process and product dimensions, especially throughout biology. The learning process is expected to develop the ability to think can high level on students. Further to the review analysis students against dimension necessary cognitive development test instrument. Therefore, this analysis is required. Through such an analysis, information on students' achievement of their cognitive dimensions can is obtained that will help teachers design more effective and efficient learning. M. Reiss et al. (1985) explains that children have difficulty relating hearts understanding the problem the concept with-concept so that requires ability reasoning comparison with or related case studies.

L.M. Neil (2010) describes the cognitive dimension ability as a learning outcome that the learning outcome can be used as a settlement in the facts contained in the study already underway, with their knowledge of the learning outcomes that are known by the teacher can be used to answer the expectations that will be known by the students so that they can be used as an initial step in developing a learning tool to build new knowledge to students through learning. A. Majid (2014) describes the assessment of learning outcomes on cognitive aspects can be seen by the results of tests carried out on students that are tailored to the learning objectives that have been made. So the success in the learning process can be seen by the achievements of the tests carried out to the students. Through tests conducted to students, teachers can provide feedback on the acquisition of student learning outcomes. E.P. Widoyoko (2014) describes the use of tests to determine student learning outcomes can provide an overview of existing information on an object or the ability of students after going through the learning process of students. A learning outcome is a learning achievement that has been obtained through a series of the learning process.

Referring to Indonesian Qualification Framework (IQF) and the revised Bloom's taxonomy (Anderson and Krathwohl, 2001), learning outcome as a learning achievement consists of some aspects, part of them are cognitive process skills and knowledge (cognitive product) mastery. By using the competencies-based curriculum, the Indonesian government has supported students to attain the learning achievement.

After the implementation of competencies-based curriculum, there is still limited comprehensive information on Indonesian secondary school students' acquisition of process skills and cognitive product, particularly in Biology. Regarding this situation, it is important to analyze how Indonesian students acquire process skills and cognitive product. In order to get a comprehensive result, such analysis needs to be conducted in regions with various types or levels of schools, teachers, and students. For the purpose of this analysis, a comprehensive, valid, and reliable instrument needs to be developed.

Development of a test instrument is needed in order to obtain optimum results in instructional processes related to students' cognitive process and product dimensions (Sumintono & Widhiarso, 2015). This is because no such test instrument has been used to analyze these students' cognitive dimensions.

Rasch (Bond & Fox, 2015) invention in the field of psychometric, wrote that a person having a greater ability than another person should have the greater probability of solving any item of the type in question, and similarly, one item being more difficult, than another means that for any person the probability of solving the second item is the greater one.

That is, individuals who have abilities better / higher compared with other individuals will have a better chance Great to answer the item correctly, and similarly, one difficult item will cause individuals opportunities to answer becomes lower. It can simply be understood that, the opportunity to be able to resolve one problem correctly depending on the ratio between the capacity of people and the level of difficulty about.

Rasch modeling also opens the information that the raw scores can not be used as a reference in estimating the ability of individuals or groups because basically raw scores do not match with the rules of measurement to be used as reference measurement results. Rasch modeling provides an understanding that, just raw scores a label-shaped figure. So, Rasch creates measurement scale which shows the same distance (equal interval) and linear. Next Rasch models have become one of the models in the measurement used in the field of education. The Rasch Model involves a degree difficulty, which is said to be similar to the model that basically IRT 1PL indeed the emphasis on the level of difficulty. However, B. Sumintono & W. Widhiarso (2015) stated that Rasch models and IRT models 1 PL has a difference, so Rasch and IRT can be seen as a model of different measurements. Although in principle between IRT and Rasch depart from the same thing, namely improvers

The weakness of classical test theory. Fundamentally, the validity of the evidence shows that instruments used can give an accurate picture the variables to be addressed in accordance with the purpose of research. S. Azwar (2015) explains that validity came from words that have meaning how far the accuracy of a test or scale the measurement function. Measurement is said to have a high validity if generate data that accurately provides an overview the measured variables as desired by destination such measurements. Then, D. Mardapi (2008) explains that, validity the evidence and theoretical support to the interpretation of test scores in accordance the intended use of the test. Therefore, the validity of the most basic fundamental in developing and evaluating a test. The validation process includes collecting evidence to show The scientific basis of interpretation of the score as planned. Furthermore, B. Subali (2012) argues that "The problems encountered in meeting the construct validity in cognitive domain is not only limited to the item complies with indicators of achievement of competencies. The fundamental issue is whether the number of competencies measured is in the one-dimensional". Based on such understanding, it is understood that the construct validation related construct of items developed, customized with the competencies to be known.

In simple terms, the reliability is understood as constancy or consistency of a measuring instrument. Definition of reliability associated with consistency. B. Subali (2012) explains that a tool otherwise reliable measure/reliable if it gives the same result at many times repeatability of measurements. More clearly D. Mardapi (2008) explains that the reliability or reliability is a coefficient that indicates the level of regularity or consistency the measurement results of a test. Consistency relates to the level error results in the form of a test score. The tests used in places with the same purpose, such as the achievement test, the results in the form of a score must be comparable between places. Result These tests must also be compared across time to find out the development of learning outcomes are achieved.

The difficulty level, Item difficulty index is said to be good if more than -2.0 or less of 2.0 which can be expressed by (-2.0> b <2.0). the level of difficulty. This can be searched using the QUEST program. Item Characteristic Curve (Item Characteristic Curve, ICC) The characteristics of the item indicated by the item characteristic curve (ICC). Item characteristic curve provides information about the relationship chance to answer correctly with the ability of learners. B. Subali, (2012) points out, Item Characteristic Curve (ICC) will form a curve horizontal (flat) when the magnitude INFIT e MNSQ for items or more logit of the unit 1.30, or less than 0.77 unit logit with an average of 1.0. When the value >1.30 consequently form a curve platykurtic (curve too blunt) and when <0.77 would be a too leptokurtic curve (curve too pointy). ICC curves obtained with the help of MG Bilog program for data dichotomous and Parscale for data polytomous. Functions of Information and SEM Function item information (item information function) basically, will produce grain information which matches the model. H. Retnawati (2016) explains that in item response theory, known as the value of the function information. Function item information (item information function) is a method to explain the strength of an item on the test device, elections test items, and comparing several test devices. function information. With function item information is known to the items which match models that assist in the selection test items. Hambleton & Swamithan (1985) explains that "the item response theory analog of the score reliability and the standard error of measurement is the test information function ". That is, in theory, item response, reliability can be known through SEM graphs on the function information.

Thus, on the basis of the descriptions presented above, the need is felt to study the dimensions of students' cognitive process and product in biology based on L.R. Anderson and D.R. Krathwohl (2001). As an initial step, the development of this test instrument is carried out in a limited number of regions in Indonesia. This is done by taking regional characteristics into consideration. Results of such analysis are expected to improve evaluation of classroom instructional material, in a narrow scope, and that of education policies, in a broader scope.

Material and method

The research is a research and development (R&D), to develop test instruments. The R&D refers and accommodates the Oriondo-Wilson's test development method (Oriondo & Dallo-Antonio, 2008; Wilson, 2005). The developed test items cover four topics on the biology of grade 11 senior high school student, second semester, there were the excretory system, coordination system, reproductive system, and the immune system.

The topics for the instruments are related to the following basic competencies: (1) to analyze the relationship between network structure constituent organs of the excretory system and link it with the process of excretion in order to be able to explain the mechanism as well as malfunctioning that may occur in the human excretory system through the study of literature, observation, experimentation, and simulation; (2) to analyze the relationship between network structure constituent organs of the coordination system and associate it with the coordination process so as to explain the role of the nervous and hormonal mechanisms of coordination and regulation as well as malfunctioning that may occur in the human coordination system through the study of literature, observation, experimentation, and simulation; (3) to analyze the relationship between network structure constituents with reproductive organ functions in the process of human reproduction through the study of literature, observation, experimentation, and simulation; and (4) to apply the principles of the understanding of the immune system to improve the quality of human life through immunization programs to maintain physiological processes in the body.

The instrument of this study was in the form of objective written test. This objective written test comprised multiple choice items and objective descriptive tests, i.e. multiple choice items in which students were asked to write their explanation. In order to measure students' metacognitive knowledge (i.e. cognitive product (K4), the present study modified the existing instrument which was developed by developed by A. Panaoura & G. Philippou (2006) and Paidi (2009). This modified instrument was in the form of Metacognitive Awareness Inventory that consisted of 29 items.

The multiple choices test consisted of 60 items for measuring cognitive process C1 to C5. The descriptive objective test consisted of 6 items which were used to measure cognitive process C6. These 60 multiple choice items were distributed into 2 booklets that each consisted of 30 items. However, at the end of item validation process in the test instrument developing, 3 items were excluded from the booklets because they were not valid.

The test covered four main topics in the second semester of Grade 11, i.e. excretory system, coordination system, reproductive system, and the immune system. These four topics were considered to be relevant for measuring students' process skills and cognitive product because these topics involved various activities, issues, and also cognitive complexities.

The each instrument test item is to measure SHS student's ability in process and product of cognitive simultaneity. The each instrument shows intersection between and cognitive (C) product (K or knowledge) and process dimensions C1K1, C2K2... until C6K4.

The research modified the instrument item development procedure from L.L. Oriondo & E.M. Dallo-Antonio (2008). The procedure consists of (1) instrument item drafting, and (2) field testing. The item drafting was be done via (a) determining aspects of competence to test, (b) determining a relevant biology subject matter(s), (c) developing table of instrument specification, (d) constructing the instrument items based on the principles of cognitive dimensions L.R. Anderson and D.R. Krathwohl (2001), (e) composing criteria for scoring, (f) reviewing the instrument item(s), and (g) revising (Istiyono, 2014).

The field testing was conducted in four districts in Indonesia, as regions sample, namely: Padang (West Sumatra), South Jakarta, Madiun (East Java), and Tenggarong (East Kalimantan). The sampling technique to take schools sample was conducted using purposive sampling in each district, three senior high schools (SHS) were selected representing high, medium, and low level of preference as seen from the results of the national examination in the previous year. The sample size of the schools to be used in research refer to Donald Ary (Furchan, 2011) explaining that the sample size for a descriptive study between 10% - 20% of the population. All the senior high schools were public schools or state senior high schools (SSHS). The schools were (1) SSHS 1 Padang, SSHS 2 Padang, and SSHS 15 Padang; (2) SSHS 8 Jakarta, SSHS 55 Jakarta, and SSHS 97 Jakarta; (3) SSHS Nglames, SSHS 1 Mejayan, and SSHS 2 Mejayan; and (4) SSHS 1 Tenggarong, SSHS 2 Tenggarong, and SSHS 3 Tenggarong. From the 12 schools, a total of 1.126 students were obtained from the research respondents.

The field testing data was be analyzed descriptively using QUEST, BILOG, and Parscale to find-out validity and reliability of the instrument items, including the goodness of fit, difficulty index, reliability estimate, item characteristic curve (ICC), Standard Error of Measurement (SEM), and standard error of measurement (SEM).

For logical validation, instruments were subjected to expert judgments concerning the aspects of material, construction, and language. Based on the inputs from the experts, the instruments of the test were finalized to be ready for piloting (field testing).

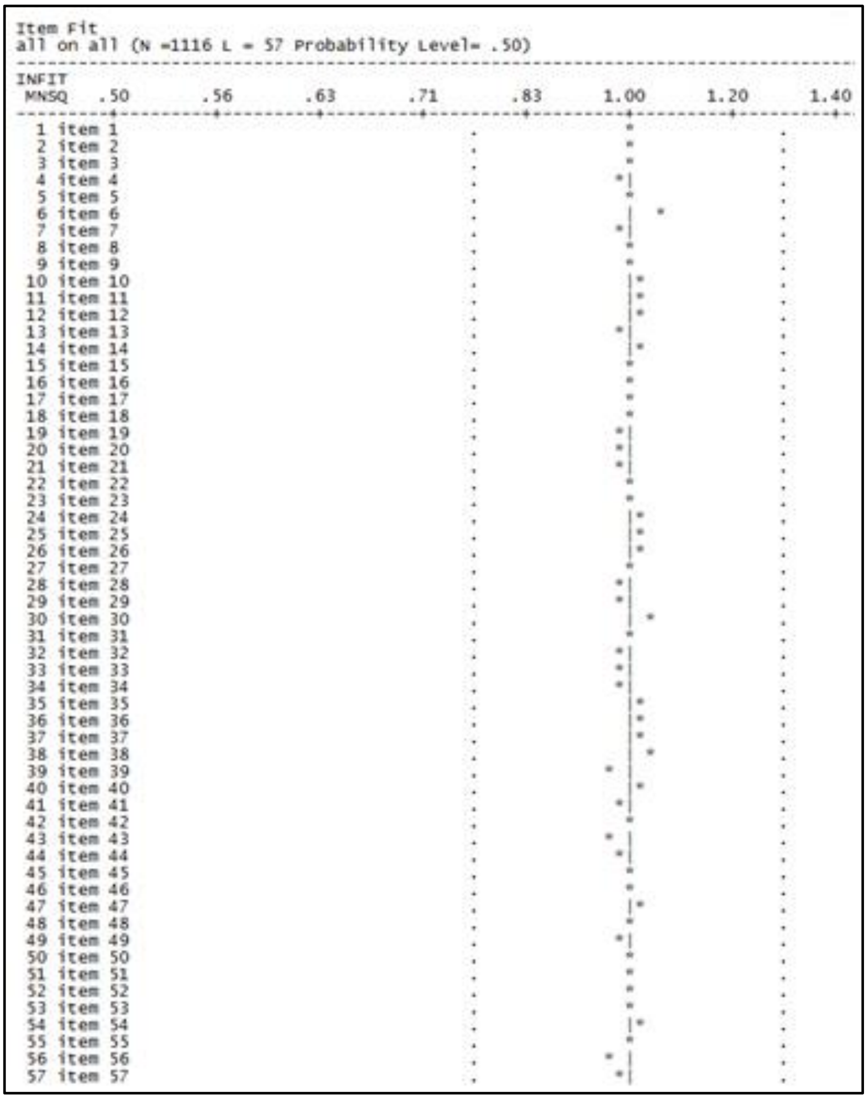
In this study, data were analyzed and interpreted using item response theory or modern test. The weakness of item response theory can be solved with the theory item response (IRT). Basically, IRT has three types of models measurement, namely (a) the model 1PL (parameter logistic) involving item difficulty, (b) 2PL models involving this level of difficulty and different power point, (c) models involving 3PL item difficulty, grain and guessing different power (guesses). Modern test theory is known as item response theory (item response theory) trying to overcome the weaknesses that are owned classical test theory. D. Mardapi (2008) argues that the theory raised by the Lord in 1952, known as the theory of test scores. Furthermore, Birnbaum develops the statistical basis for models of item response theory in 1957. Furthermore, the theory developed by other researchers. One of the researchers who developed the appropriate measurement model with item response theory is George Rasch in 1960. Rasch stated that he developed the theory refers to the model probabilistic. The purpose of probabilistic models developed by Rasch obtained through analysis of the raw scores on exam results learners primary school age (Mardapi, 2008).

Results and discussion

Instruments test for the cognitive dimension of grade 11 senior high school student has been validated to one related expert prior to trial. The test is done in order to determine each item better readability test items. The instruments test that has been developed consisting of a set of multiple choice questions amounted to 57 grains consisting of C1-C5 aspects and 6 questions description consisting of C6 aspect. Based on the results of the validation tests conducted by the instrument of measurement experts and subject matter experts further revision and refinement of the items. Event subsequent revisions to the instrument used for the test phase. Results input based on validation by subject matter experts as well as experts on the measurement test instrument with respect to the suitability of the material with a concept map drawn up as well as the appearance of images, graphs/tables are displayed, whereas for expert measurements are input with respect to fitness for purpose, research for each dimensional aspects of cognitive processes and knowledge.

Program QUEST is used to find the validity and reliability measures of the test. The test validity of the Rasch model can be seen from the item fit to the model. Using the 5% error limit, an item is said to fit the model if the INFIT MNSQ score is between 0.77 and 1.30 and the INFIT *t* is between -2.0 and 2.0. In addition, item curve characteristics (ICC) and information function graphs are presented using the Bilog and Parscale programs. Meanwhile, for the metacognitive non-test, validity and reliability measures are obtained by way of the SPSS program.

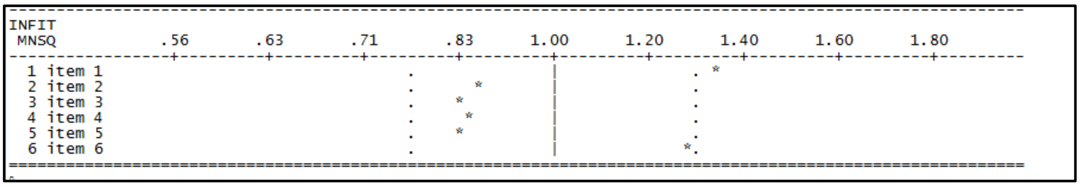
The Rasch analyses show that 57 test items fit the model since they have the score criteria for the INFIT MNQS between 0.77 and 1.30 as shown in Figure 1. This shows that each item in the instrument is empirically valid for measuring students' competencies in the cognitive process and product dimensions.



```
Item Fit
all on all (N =1116 L = 57 Probability Level= .50)
-----------------------------------------------------------------------------------
INFIT
 MNSQ   .50       .56       .63       .71       .83      1.00      1.20      1.40
----------+---------+---------+---------+---------+---------+---------+---------+--
  1 item 1                                  .               *              .
  2 item 2                                  .               *              .
  3 item 3                                  .               *              .
  4 item 4                                  .              *|              .
  5 item 5                                  .               *              .
  6 item 6                                  .               |  *           .
  7 item 7                                  .              *|              .
  8 item 8                                  .               *              .
  9 item 9                                  .               *              .
 10 item 10                                 .               |*             .
 11 item 11                                 .               |*             .
 12 item 12                                 .               |*             .
 13 item 13                                 .              *|              .
 14 item 14                                 .               |*             .
 15 item 15                                 .               *              .
 16 item 16                                 .               *              .
 17 item 17                                 .               *              .
 18 item 18                                 .               *              .
 19 item 19                                 .              *|              .
 20 item 20                                 .              *|              .
 21 item 21                                 .              *|              .
 22 item 22                                 .               *              .
 23 item 23                                 .               *              .
 24 item 24                                 .               |*             .
 25 item 25                                 .               |*             .
 26 item 26                                 .               |*             .
 27 item 27                                 .               *              .
 28 item 28                                 .              *|              .
 29 item 29                                 .              *|              .
 30 item 30                                 .               | *            .
 31 item 31                                 .               *              .
 32 item 32                                 .              *|              .
 33 item 33                                 .              *|              .
 34 item 34                                 .              *|              .
 35 item 35                                 .               |*             .
 36 item 36                                 .               |*             .
 37 item 37                                 .               |*             .
 38 item 38                                 .               | *            .
 39 item 39                                 .             * |              .
 40 item 40                                 .               |*             .
 41 item 41                                 .              *|              .
 42 item 42                                 .               *              .
 43 item 43                                 .             * |              .
 44 item 44                                 .              *|              .
 45 item 45                                 .               *              .
 46 item 46                                 .               *              .
 47 item 47                                 .               |*             .
 48 item 48                                 .               *              .
 49 item 49                                 .              *|              .
 50 item 50                                 .               *              .
 51 item 51                                 .               *              .
 52 item 52                                 .               *              .
 53 item 53                                 .               *              .
 54 item 54                                 .               |*             .
 55 item 55                                 .               *              .
 56 item 56                                 .             * |              .
 57 item 57                                 .              *|              .
```

Figure 1. Diagram of INFIT MNSQ for the multiple choice items in the tryout phase

From Figure 1, it can be seen that the difficulty level of the instrument ranges between -2 and +2, indicating that the test is good for use. Results of analyses for the objective description test-type test items can be seen in Figure 2.



```
------------------------------------------------------------------------------------------------------------
INFIT
 MNSQ           .56       .63       .71       .83      1.00      1.20      1.40      1.60      1.80
-----------------+---------+---------+---------+---------+---------+---------+---------+---------+----------
  1 item 1                                .              |              . *
  2 item 2                                .      *       |              .
  3 item 3                                .    *         |              .
  4 item 4                                .     *        |              .
  5 item 5                                .    *         |              .
  6 item 6                                .              |             *.
============================================================================================================
```

Figure 2. Diagram of INFIT MNSQ for the description items

In this figure, the average score of the *INFIT MNS* is 0.99 with a standard deviation of 0.13. It can be stated that the objective description test items have a fit with the Rasch model. Meanwhile to determine that an item has a *fit* to the model is to see that the *INFIT MNSQ* is found between 0.77 and 1.30 and the INFIT *t* is between -2.0 and 2.0. Thus, it can be stated that the test items have full filled the criteria for the *goodness of fit*.

The reliability measure of the multiple-choice test items is obtained by using the QUEST program. The results of the analyses for reliability show a coefficient of 0.98.The estimation results of the reliability analyses are presented in **Table 2** below.

Table 2. Estimation Results for the Multiple-Choice Test Items

Aspect	Item estimate	Case estimate
Reliability	0.98	
The average value and standard deviation of the INFIT MNSQ	1.00 ± 0.02	1.00 ± 0.05
The average value and standard deviation of the OUTFIT MNSQ	1.01 ± 0.12	1.00` ± 0.16

For the description test item, the QUEST program gives the reliability estimate value of 0.93.The estimates are shown in Table 3 below.

Table 3. Estimation Results for the Objective description test Items

Aspect	Item estimate	Case Estimate
Reliability	0,93	
The average value and standard deviation of the INFIT MNSQ	1.00 ± 0.24	1.32 ± 0.64
The average value and standard deviation of the OUTFIT MNSQ	1.09 ± 0.39	1,09 ± 1.19

For the measures of the item difficulty levels, the QUEST gives scores as can be seen in **Figure 3**. A test item is said to be good if the difficulty index is > -2.0 or <2.0. The most difficult item is seen in the C3P aspect: cognitive process application and cognitive product procedural knowledge. The easiest item is seen in the C1P aspect: cognitive processes call and cognitive product procedural knowledge.

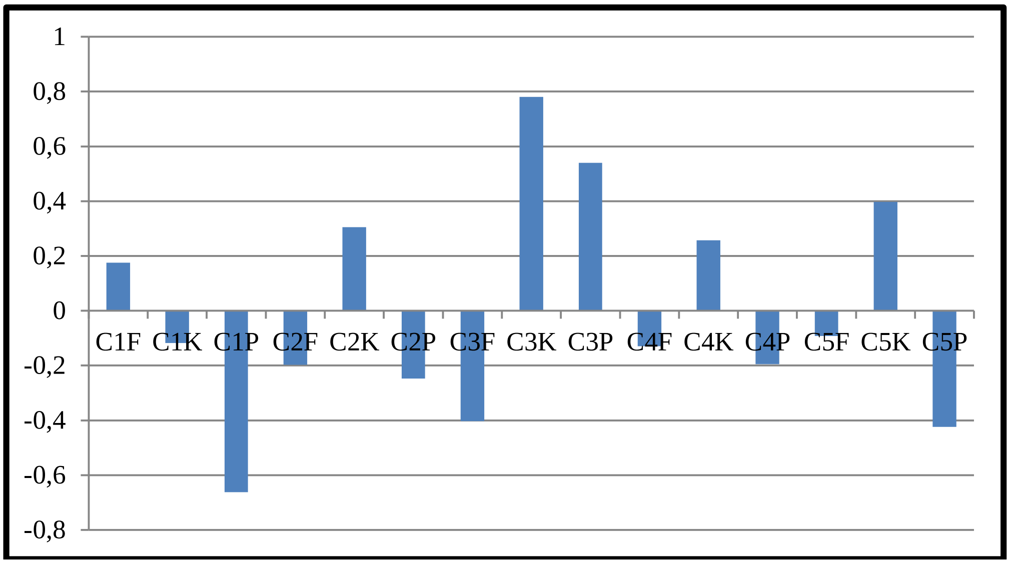


Figure 3. Distribution of difficulty levels for the multiple-choice items

Notes

C1 F: Remember-Factual knowledge
C1 K: Remember-Conceptual knowledge
C1 P: Remember-Procedural knowledge
C2 F: Understand-Factual knowledge
C2 K: Understand-Conceptual knowledge
C2 P: Understand-Procedural knowledge
C3 F: Apply-Factual knowledge
C3 K: Apply-Conceptual knowledge
C3 P: Apply-Procedural knowledge
C4 F: Analyze-Factual knowledge
C4 K: Analyze-Conceptual knowledge
C4 P: Analyze-Procedural knowledge
C5 F: Evaluate-Factual knowledge
C5 K: Evaluate-Conceptual knowledge
C5 P: Evaluate-Procedural knowledge

For the description test item, the distribution of the difficulty level of each category can be found individually. Distribution of the difficulty level can be seen in Figure 4.

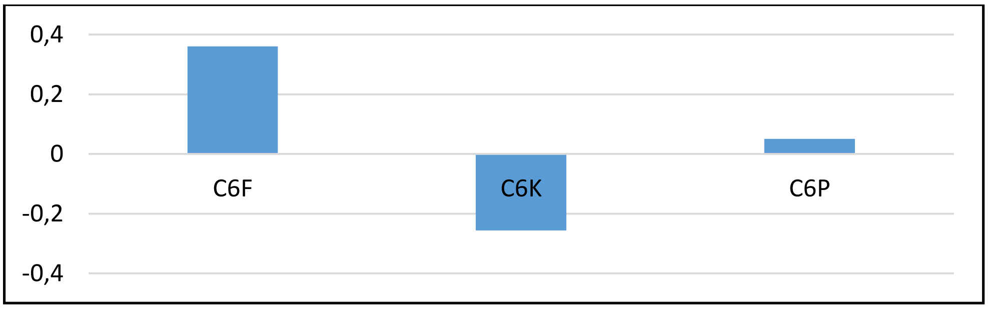


Figure 4. Distribution Difficulty Test of Description item

Notes:

C6F: Create Factual
C6K: Create Conceptual
C6P: Create Procedural

Item characteristics curve indicated with using item (ICC) The Program lead with program Bilog MG. Furthermore, based on the results of the analysis using the Bilog MG program, the item characteristic curve for each item is obtained. Figure 5 shows an example of the ICC of item number 18.

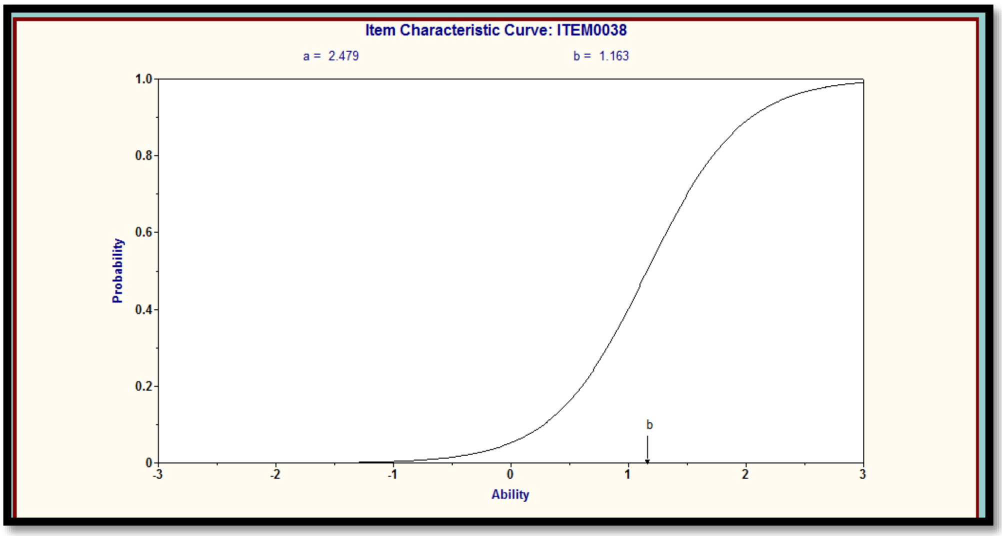


Figure 5. Item Characteristic curve for Item Number 18

From Figure 5 above, it is understood that item number 18 can be done by learners with the ability (b) or a high capacity since the peak of the curve stands at ± 1.3. Thus, it can be said abilities (b) learners who can work on these items is high or converted into a high capacity.

The characteristics of the item description on the test instrument, it is shown by the item characteristic curve (ICC). Item characteristic curve (ICC) is raised to the program description Parscale. Then, by using the Parscale program, each ICC obtains as many as 6 pieces. Figure 6 shows the ICC for item number 10, or number 4 in package II. From this figure, the ICC for item number 10 can be explained as follows: (1) a score of 1 (category 1) is mostly obtained by learners with low ability (θ = -3); (2) a score of 2 (category 2) is mostly obtained by learners with low ability (θ = -0.5); (3) a score of 3 (category 3) is mostly obtained by learners with high ability (θ = 0.9); and (4) a score of 4 (category 4) is mostly obtained by learners with high ability (θ = 3).

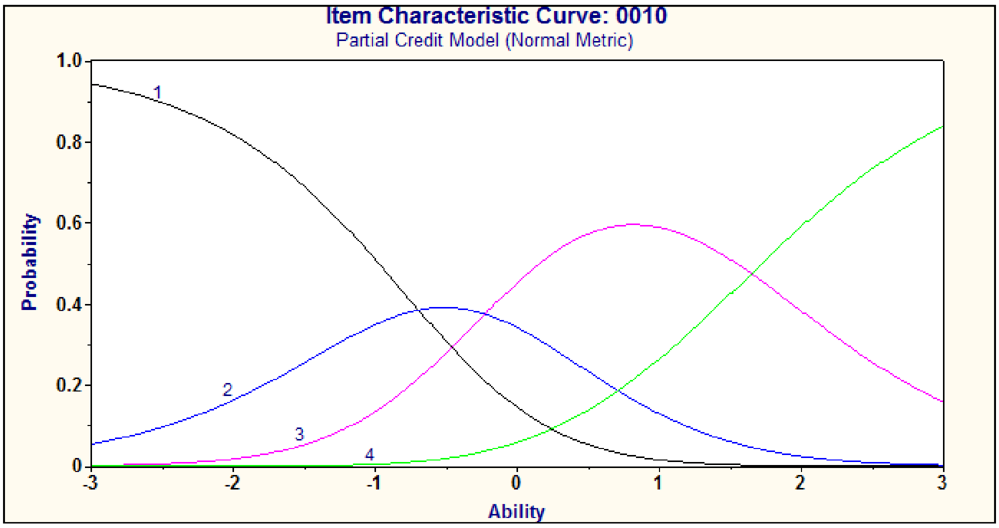


Figure 6. Curve Characteristics of Problem Description of Item Number 10

Furthermore, the graph of the function information and the MCQ SEM is presented in Figure 7 below.

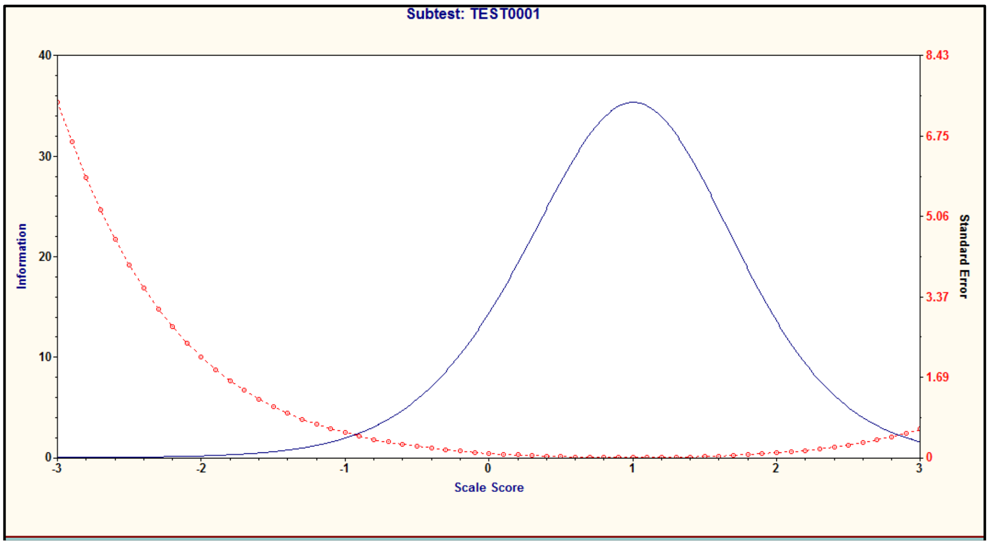


Figure 7. Function Information and Standard Error Measurement (SEM) of the Multiple Choice Questions (MCQ)

Based on this figure, it can be seen that the test description instrument is suitable for learners with low to moderate abilities, namely: -1.4 <θ <2.8. -1.0 ≤ θ ≤ 2.8. Then, information for the graph function and SEM for the described problems is presented in Figure 8.

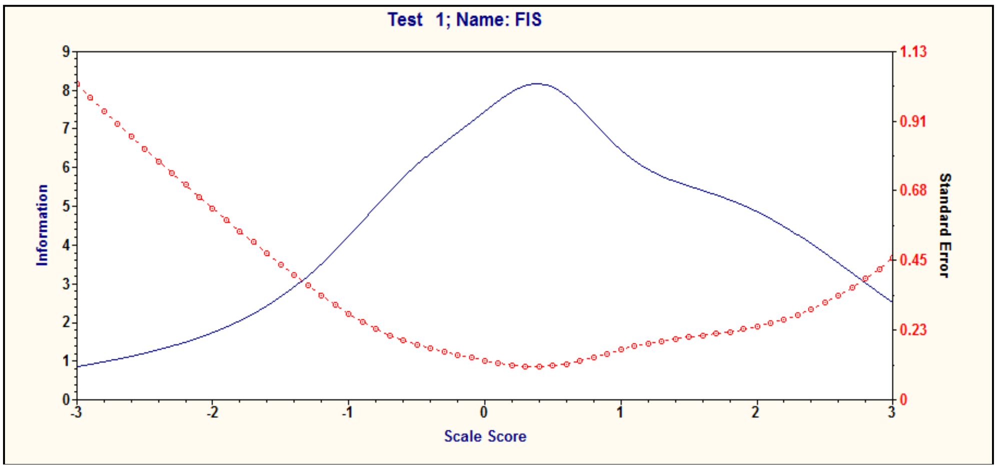
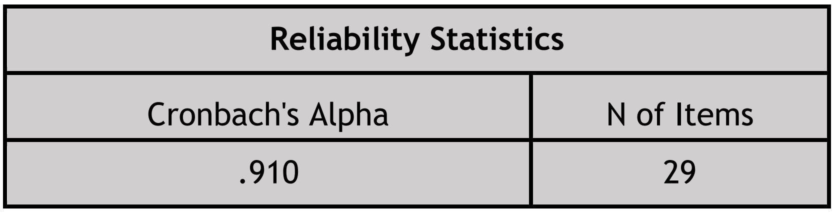


Figure 8. The Information and Standard Error Measurement (SEM) of the Description Questions

This figure shows two peaks of information, which means that there are two optimal information pieces obtained by the test, i.e. at low and high ability individuals simultaneously. Thus, it can be stated that the test dimensions of cognitive process and knowledge dimensions are appropriate for learners who have medium and high ability categories of -0.7 ≤ θ ≤ 0.8.

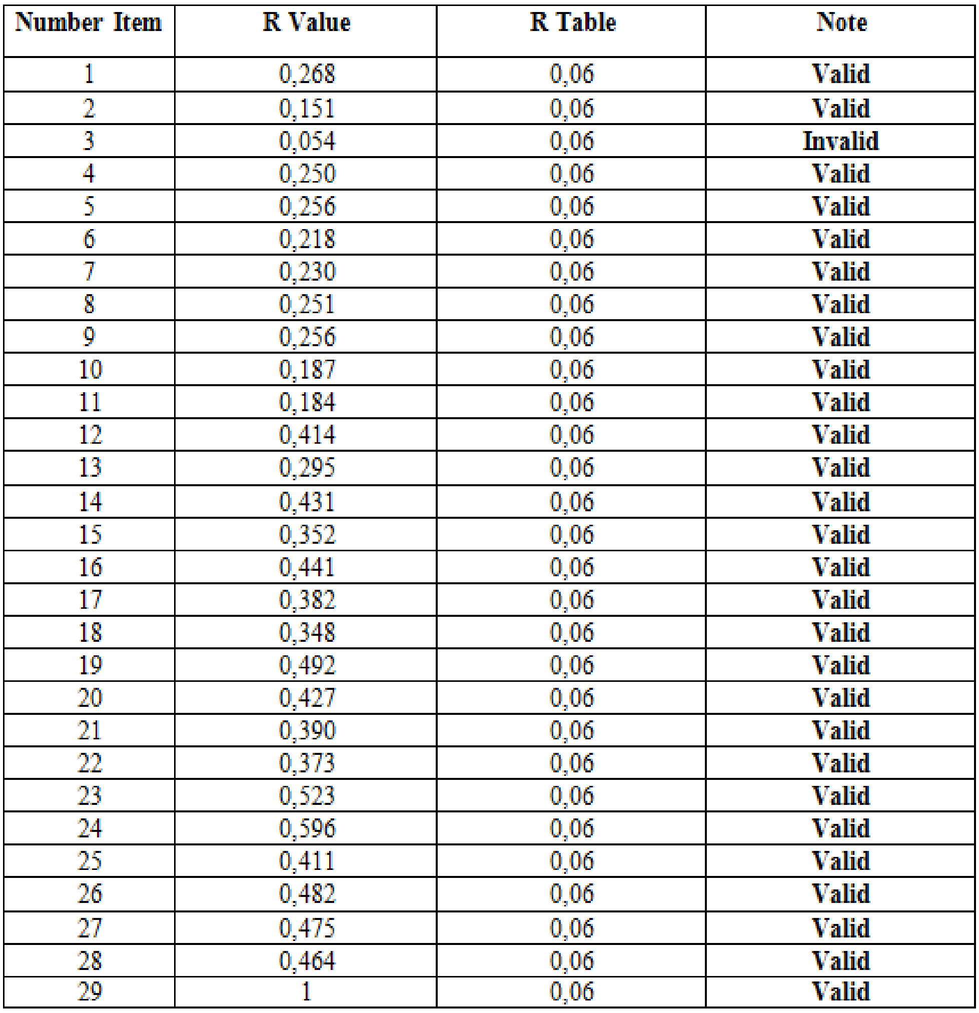
Finally, from the results of the reliability and validity analyses for the Awareness metacognitive inventory through the SPSS, the reliability of the non-test instrument can be seen in Figure 9.



Reliability Statistics	
Cronbach's Alpha	N of Items
.910	29

Figure 9. Reliability Coefficient for the Non- Test Metacognitive Instrument

From Figure 9, it can be seen that the reliability measure of the metacognitive instrument is 0.91. Meanwhile, the validity of metacognitive instrument can be seen in Figure 10.



Number Item	R Value	R Table	Note
1	0,268	0,06	**Valid**
2	0,151	0,06	**Valid**
3	0,054	0,06	**Invalid**
4	0,250	0,06	**Valid**
5	0,256	0,06	**Valid**
6	0,218	0,06	**Valid**
7	0,230	0,06	**Valid**
8	0,251	0,06	**Valid**
9	0,256	0,06	**Valid**
10	0,187	0,06	**Valid**
11	0,184	0,06	**Valid**
12	0,414	0,06	**Valid**
13	0,295	0,06	**Valid**
14	0,431	0,06	**Valid**
15	0,352	0,06	**Valid**
16	0,441	0,06	**Valid**
17	0,382	0,06	**Valid**
18	0,348	0,06	**Valid**
19	0,492	0,06	**Valid**
20	0,427	0,06	**Valid**
21	0,390	0,06	**Valid**
22	0,373	0,06	**Valid**
23	0,523	0,06	**Valid**
24	0,596	0,06	**Valid**
25	0,411	0,06	**Valid**
26	0,482	0,06	**Valid**
27	0,475	0,06	**Valid**
28	0,464	0,06	**Valid**
29	1	0,06	**Valid**

Figure 10. Results of the Validation of the Metacognitive Instrument

In this table, it can be seen that, based on the tryout results, out of the 29 items tested, 28 items are found valid, and one not valid. Thus, based on the Rasch analyses on the multiple-choice and description test questions of the tryout results, it is found that developed instrument is suitable for measuring the students' cognitive process and the product (knowledge dimensions) in biology. However, it is also found that, based on the analyses of the difficulty levels of the multiple-choice and description tests, there is some inconsistency within the items. This is because the test instrument is less able to demonstrate the hierarchy of the levels of the cognitive abilities of the learners. Therefore, the test items need to go through stages of revision so that the test instrument has a stronger power to be used for testing students' cognitive process and product dimensions in biology.

Finally, the non-test metacognitive instrument, however, is found to have a high measure of reliability and validity that meets the requirement for non-test instrument development. So it can be used to analyze students' metacognitive abilities in biology.

Conclusion

Based on the description and discussion of the research findings, the research study has produced the following results. First, the multiple-choice test has a mean and standard deviation of 1.0 and 0.0 which fits the INFIT MNSQ and the description test fits the Rasch model. Second, the INFIT MNSQ lower and upper bounds of 0.77 and 1.30 indicate that there are items that do not fit the models. Third, based on the analysis results of the item difficulty levels, items that represent aspects do not show the hierarchy of the cognitive capability dimensions. Finally, revision of some of the items is needed that will be used in various stages of implementation. On the other hand, the metacognitive instrument is found to have a high measure of reliability and validity that meets the requirement for non-test instrument development and can be used to analyze students' metacognitive abilities in biology.

Dısclosure statement

The Authors reported that no competing financial interest.

Notes on contributors

Paidi - is a faculty member in Biology Education Department, Faculty of Mathematics and Sciences, Yogyakarta, Indonesia

Djukri - Faculty of Mathematics and Sciences, Yogyakarta State University (UNY), Yogyakarta, Indonesia

Siti Yulaikah - Biology Education Graduate Program, Yogyakarta State University, Yogyakarta, Indonesia.

Dessy Alfindasari - Biology Education Graduate Program, Yogyakarta State University, Yogyakarta, Indonesia.

References

Anderson L.R, Krathwohl, D.R. (2001). *A Taxonomy for learning, teaching, and assessing*: A Revision of Bloom's Taxonomy of Educational Objectives. A Bridged Edition. New York: Longman.

Azwar, S. (2015). *Reliability and validity*. Yogyakarta: Student Library.

Bloom, B.S. (1956). *Taxonomy of educational objectives: The classification of educational goals: Handbook I, cognitive domain*. New York: Longman.

Bond & Fox. (2015). *Applying the Rasch Model: the fundamental measurement in the human sciences*. 2nd Ed. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Dettmer. (2006). New Blooms in established fields: Four Domains of Learning and Doing. *Proquest Education Journals, 28*(2), 164-178.

Furchan, A. (2011). *Pengantar penelitian dalam pendidikan) (Research introduction in education*). Pustaka Pelajar. Yogyakarta

IEA. (2011). *TIMSS & PIRLS.IEA Sites*. Accessed on October 26, 2016, from http://timssandpirls.bc.edu/data-release-2011/pdf/Overview-TIMSS-and-PIRLS-2011-Achievement.pdf.

Istiyono, E. (2014). *Measurement of high-level thinking skills of high school physics students in DIY* (Doctoral dissertation). Yogyakarta: State University of Yogyakarta.

Majid, A. 2014. *Penilaian autentik proses dan hasil belajar*. Bandung PT Remaja Rosdakarya.

Mardapi, D. (2008). *Mechanical preparation of test and nontest instruments*. Yogyakarta: Cendikia Partners Press.

Ministry of Education, Province of British Columbia. (2008). *Science And Technology 11. Integrated Resource Package 2008.* Library and Archives Canada Cataloguing in Publication Data.

Neil, L.M. (2010). *Contradictions of school reform: Educational Coast Of Standardized Testing.* (electronically version). New York: Taylor & Francis e-library

Nitko, A.J.& Brookhart, S.M. (2011). *Educational assessment of student* (6th ed). New Jersey: Pearson Education Inc.

Oriondo, L.L. & Dallo-Antonio, E.M. (2008). *Evaluation of educational outcomes*. Manila: Rex Printing Company, Inc.

Paidi. (2009). Developing of Problem-Based Instruction Materials on Biology and Metacognitive Strategy and Its Effectiveness to Metacognitive Awareness, Problem Solving Skill, and Subject Matter Mastering of High School Student in Sleman-Yogyakarta. *Doctoral dissertation*, unpublished. Malang: University of Malang.

Panaoura, A & Philippou, G. (2006). *The measurement of young pupils´ metacognitive ability in mathematics: The Case of Self-Representation and Self-Evaluation*. Department of Education, University of Cyprus. Direct access: http://cerme4.crm.es/Papers%20definitius/2/panaoura.philippou.pdf.

Reiss, M., Behr, M., Lesh, R., & Post, T. (1985). Cognitive Processes And Products in Proportional Reasoning. In L. Streefland (Ed.), *Proceedings of the Ninth International Conference for the Psychology of Mathematics Education* (pp. 352-356). Noordwijkerhout (Utrecht), Holland: PME.

Retnawati, H. (2016). *Validity, reliability, and characteristics of the grain*. Yogyakarta: Parama Publishing.

Subali, B. (2012). *Test measurement science process skills divergent patterns of biological subjects.*Yogyakarta State University.

Suciati. (2015). *Memahami hakikat dan karakteristik pembelajaran biologi dalam upaya menjawab tantangan Abad X1 serta optimalisasi impelementasi Kurikulum 2013*. Surakarta: UNS.

Sumintono, B. & Widhiarso, W. (2015). *Rasch modelling applications in educational assessment*. Cimahi: Trim Komunikata.

Tutkun O.F, Guzel D, Koroğlu M, Ilhan H. (2012). Bloom's Revised Taxonomy and Critics on It. *TOJCE: The on line Journal of Counselling and Education,1*(3), 253-269.

Widodo, A. (2006). *Revisi taksonomi boom dan pengembangan butir soal.* Bandung. *Buletin Puspendik, 3*(2), 18-29.

Widoyoko, E.P. (2014). *Schools learning outcomes assessment.* Yogyakarta. Pustaka Pelajar

Wilson, M. (2005).*Constructing measures: An item response modeling approach.* Mahwah: Lawrence Erlbaum Associates, Inc. Publishers.

Wu, X.N, Wu, X. & Wang, W. (2016). How Do Cognitive and Affective Trust Impact Process Outcome Interaction. *Social Behavior and Personality*, *44*(8), 153-174.

Yu-Mei Lin & Pei-Chen Lee. (2013). *The practice of business's teacher teaching*: Perspective from critical thinking. Taipei: China Institute of Technology.