



JOURNAL
OF BALTIC
SCIENCE
EDUCATION

ISSN 1648-3898 /Print/

ISSN 2538-7138 /Online/

Abstract. *The aim of this research was to assess the classification of science test items of TIMSS grade 8 based on higher order thinking skills (HOTS) and determine whether those classified-science test items can be an assessment tool in science class.*

Sixteen sample test items of HOTS were chosen from 37 reasoning items of TIMSS 1999, 2003, and 2011; which were 6 of analysing, 6 of evaluating, and 4 of creating. The selected items were tested to 410 ninth grade students in 14 public schools in Jember, Indonesia. Data were analysed by using point-biserial correlation to measure the index of discrimination and degree of difficulty at items of each level of HOTS test.

The result revealed that the point-biserial index of discrimination for each item was higher than 0.25. The degree of difficulty of analysing, evaluating and creating test items exhibited a similar trend, which was in good range. Each test item has significant validity. Whilst reliability analysis showed that each test item was acceptable and indicating a high level of internal consistency. In conclusion, the classified science test items of TIMSS are good to use as assessment tools to measure HOTS of students in science class.

Keywords: *higher order thinking skills, point biserial correlation, science test items.*

Anjar Putro Utomo, Erlia Narulita
University of Jember, Indonesia
Kinya Shimizu
Hiroshima University, Japan

DIVERSIFICATION OF REASONING SCIENCE TEST ITEMS OF TIMSS GRADE 8 BASED ON HIGHER ORDER THINKING SKILLS: A CASE STUDY OF INDONESIAN STUDENTS

**Anjar Putro Utomo,
Erlia Narulita,
Kinya Shimizu**

Introduction

Indonesian education system is immense and diverse. With over 60 million students from primary to secondary level and almost 4 million teachers in some 340000 private and public schools, it is the third largest education system in the Asia region and the fourth largest in the world (behind only the People's Republic of China, India, and the United States) (Organisation for Economic Cooperation Development, 2015). Consequently, those conditions should give beneficiary for an education system in Indonesia becoming better or at least in same quality with the neighbour Asian countries such as Singapore, Malaysia, and Thailand. Because there are many human resources who will make more chance to face global competition. In fact, the results of the mean score of national examination (UN) of academic year 2014/2015 is 59.88 with the highest score is 100 and the lowest score is 2.5 (Pusat Kurikulum dan Perbukuan, 2016). While a standard of the good national examination score is 70-85 (Ministry of Education and Culture, 2015). Thus, there is a big gap between mean score and the good standard score of national examination. Additionally, the range of students who got the highest score and the lowest score are very far.

These results show that Indonesia remains needs to improve their education quality. The quality of national education one of which can be seen from the output quality, the views of quality graduates recognized at the national, regional, and international. In this context, the national education, which has quality graduates, is a necessity because without producing quality graduates, the education program is not seen as a human resource investment to improve the nation's competitiveness, but it is seen as a waste in terms of costs, energy, and time. Especially in skills quality of Indonesia students is higher order thinking skills. Because one of Indonesian science educational goals is to improve student quality through developing higher



order thinking skills (HOTS) since the early age as basic skills for daily life, apart from the academic achievement in the schools. Higher order thinking involves a variety of thinking processes applied to complex situations and having multiple variables (King, Goodson & Rohani, 2015); and it is important in science education to make the relationships between evidence and explanations. It is why this issue remains to become a substantial concern of Indonesia government till now.

One of the ways to know whether Indonesia's learning has been directing to the formation of higher order thinking skills, the Ministry of Education and Culture send some representation of Junior High School/ Islamic Junior High School (SMP/MTs) students of Indonesia in the international study, called TIMSS (Trends in International Mathematics and Science Study). There are two domains that are being tested in TIMSS assessments, namely content and cognitive domains. For the tested cognitive domain, it covers knowing, applying, and reasoning (Mullis & Martin, 2013). The concerning fact is Indonesian students always got low rank and score at TIMSS studies. TIMSS result revealed that Indonesian students performed lower than the international average benchmarking in all cognitive domains, mainly in reasoning domain. An average scale score of Indonesian students in 2011 result is 406 while International average benchmark is 500. In this time, Indonesia was rank 40 out of 42 or third rank from the bottom. Especially in reasoning domain score was 20% while the international average was 33% (Martin, Mullis, Foy, & Stanco, 2012).

Indonesia student's achievement was getting a jump in 2011 result from 2007 result when comparing with the same country especially neighbour countries such as Malaysia, Thailand, Singapore, China, and Japan. This result showed that more than 95% Indonesia students can do up to intermediate level only, while 40% of the other country's students such as Singapore can reach high and advance level. It is a really alarming condition for science education in Indonesia. Almost nothing Indonesia achieved within 4 years instead Indonesia's achievement got a decrease. Furthermore, Indonesia has a huge gap with the neighbour countries like Malaysia, Thailand even less Singapore (Ministry of Education and Culture, 2013). Thus, Indonesia has a heavy duty to solve this problem and escalate student's achievement in TIMSS.

These results provide evidence of Indonesian students' difficulty in solving science tasks, which involve analysing, evaluating and creating, key aspects of HOTS that include in reasoning items of TIMSS. Consider at TIMSS assessment views, among the factors that caused the failures are the test items unfamiliar in Indonesia and students practice HOTS test items rarely (Wasis, 2014). Assessment in Indonesia mostly measures the knowledge dimension until level C3 (applying) especially in elementary and junior high school and non-linear with TIMSS cognitive domain (Wasis, 2014). Whereas, assessment can be implemented to help the students on improving their HOTS (Van den Berg, 2008). This is supported by the other argument, that HOT's question can encourage the students to think deeply about the lesson (Barnett & Francis, 2012). Since Indonesia uses the results of National Examination (UN) such as school rankings, the number of unsuccessful learners, and absorption abilities of the subjects tested in the UN, and other test results organized school (such as daily test, blocks examination, midterm examination, final exams / summative, as well as class promotion test and school exams) only as a teaching and learning feedback. Therefore, identification of TIMSS science test items in reasoning level become the level of HOTS and using it as students' HOTS assessment tool in Indonesia are a very important becoming reflection for science teaching-learning and training student's HOTS.

Focus and Aim of Research

Based on background above, Indonesia has just received reasoning result of TIMSS tests while Indonesia used HOTS level by Anderson and Krathwohl (2001) as a goal of learning. Thus, classification reasoning items of TIMSS to HOTS test items really need to give obvious improvement feedback to facing the next TIMSS test. Besides, there has been no research yet that classified reasoning science items of TIMSS based on higher order thinking skills mainly in Indonesia. The aim of this research was to identify science tests in TIMSS reasoning items into three levels of HOTS which is propitious as an assessment tools to measure HOTS of Indonesian students. This research is expected to have beneficiary for education in Indonesia especially, which are:

1. Teacher gets mapping of HOTS level in reasoning science items of TIMSS.
2. Teacher can use identified TIMSS items as assessment tool in science class.
3. Teacher gets reflection of identified TIMSS items test result of students.



Methodology of Research

General Background and Design

This research is a developmental research of science test items of TIMSS based on HOTS. HOTS test level of science in reasoning cognitive domain of TIMSS was classified; and these were determined whether or not to be used as an assessment tool as well; and then the result of this research was compared to Indonesia result and international benchmark in TIMSS 2011. The research was conducted at the beginning of even semester of academic year 2015-2016 over a period of approximately 4 weeks.

Sample Selection

Participants of the research were 410 ninth grade students from forty lower secondary schools in the area of administrative of Jember, Indonesia. One class of ninth grade was selected from each of those schools. Ten schools consisted of twenty-ninth students and four schools consisted of thirty students in each class, respectively. The ninth-grade students were selected as participants due to them already studied eight grade science topics.

The ethical guidelines for science educational research were followed during this research. The research proposal was submitted to the committees in each school, and the approval was obtained from the school within a week. The aim of the testing was informed to the students before conducting the research. The collected data were treated as confidential.

Instrument and Procedures

The 37 reasoning items of TIMSS 1999, 2003, and 2011 were organized; then these were classified into analysing, evaluating, and creating level of HOTS by a table of HOTS's component term of Krathwohl (2002, Table 1).

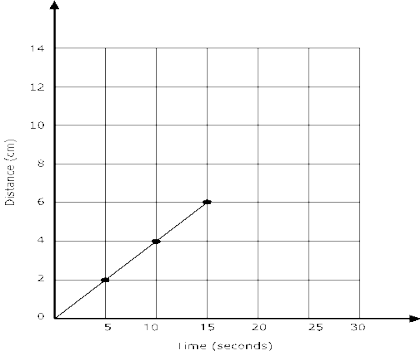

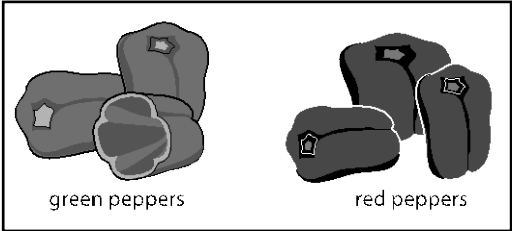
Table 1. HOTS's component term.

Analysing	Evaluating	Creating
Break material into its constituent parts and determining how the parts are related to each other and to an overall structure. For example: comparing, contrasting, differentiating, separating, identifying, relating, illustrating, etc.	Make judgments based on criteria and standards. For instance: Comparing, concluding, criticizing, critiquing, describing, measuring, assessing, explaining, interpreting, justifying, summarizing, etc.	Put elements together to form a coherent or functional whole; that is, reorganizing elements into a new pattern or structure. For instance: Categorizing, combining, collaborating, compiling, designing, generating, modifying, organizing, reorganizing, relating, reconstructing, etc.

After that, two items were randomly selected from each science subject (life science, physics, and chemistry) and level of HOTS item. Total sample test items of HOTS were 16; which were 6 of analysing, 6 of evaluating, and 4 of creating, respectively. Some examples of test items were presented in Table 2. Then those items were translated into Indonesian language. The selected items were tested to 410 of ninth grade students in 14 lower secondary public schools in Jember, East Java-Indonesia.



Table 2. Sample of test items used in this research.

Science Items	Level of HOTS	Reason
<p>The graph shows the progress made by a beetle moving along a straight line.</p> 	Evaluating	Students have to measure the correct answer base on the theory.
<p>If the beetle keeps moving at the same speed, how long will it take to travel 10 cm?</p> <p>a. 4 seconds b. 6 seconds c. 20 seconds d. 25 seconds</p>		
	Analysing	Students have to find what effect one part of ecosystem to the other part in an ecosystem
<p>The figure above shows a community consisting of mice, snakes and wheat plants. What would happen to this community if people killed the snakes?</p>		
<p>Kayra and Emre are studying plants. They have learned that characteristics such as the height of plants and the color of fruit are inherited. They are looking at some green and red peppers.</p>	Creating	Students have to arrange an investigation process to determine correct thing.
		
<p>Kayra thinks they are different kinds of peppers, because they are different colours. Emre thinks that they are the same type of pepper, and red peppers are red because they have been left on the plant longer and have ripened. Describe how you could set up an investigation to decide whether Kayra or Emre is correct.</p>		



Data Analysis

Analysis of Point-Biserial Correlation

Data were got from scoring test result of 410 students. In order to analyse point biserial correlation coefficient value, there are two kinds of category answers only that are the correct answer with score 1 and incorrect with score 0. The scoring test result was organized and calculated that data by using the formula below in excel to know index value of point-biserial correlation.

$$\rho_{pbis} = \frac{(\mu_+ - \mu_x)}{\sigma_x} \sqrt{p/q}$$

μ_+ = the mean total test score for those who answered the item correctly,

μ_x = the mean total test score for the entire group,

σ_x = standard deviation,

p = the proportion of students answer correctly, and

$q = (1-p)$

(Crocker & Algina, 1986)

Based on Ebel & Frisbie (1991), there are several following rules of thumb for determining the quality of items with respect to their point-biserial correlation coefficient index.

The Degree of difficulty

The formula of the degree of difficulty as follows.

$$p = \frac{c}{n} \times 100$$

p = the proportion of student answered correctly

c = the number of students who selected the correct answer, and

n = the total number of respondents

(Hotiu, 2006)

The "p" in the formula of the degree of difficulty is same as "p" in the formula of point-biserial correlation. It is strengthened by Crocker & Algina (1986) who stated when an item is dichotomously score, the mean item score corresponds to the proportion of examinees who answer the item correctly. This proportion is usually denoted as "p" and is called the item difficulty. It means that there is a relation between the degree of difficulty and point-biserial correlation value. So, when point-biserial correlation value is counted, automatically the value of the degree of difficulty can be known. The p-value is usually represented in decimal number after converted from percentage number. There are several categories to determine the value of the degree of difficulty (Frank, 1962 in Suruchi & Rana, 2014).

Validity

In relation with this research, content validity was measured to know the validity of HOTS's test items. This case also strengthened by the argument that content validity is the degree of correspondence between the contents of the test and the logical and curricular domains intended to be measured. The analogy items have been designed and constructed to measure knowledge, skills, and abilities considered necessary for success in graduate school (Miller Analogies Test, 2012). Validity was analysed by using Pearson correlation coefficient in SPSS. And the value of each item can be significant if the value is more than r -table (Crocker & Algina, 1986). N means sample in this research was 410 students. Since the significantly used was at the .01 level, thus r -table value based on amount sample of this research was .128. Then, the items should be significant when the Pearson correlation coefficient value is more than .128.



Reliability

This research measured internal consistency reliability is comparing the variance of each item to total test variance analysed by Cronbach's alpha in SPSS. Internal consistency reliability refers to the homogeneity of items intended to measure the same quantity (e.g., the active/reflective preference) that is the extent to which responses to the items are correlated. Cronbach's coefficient alpha, an average of all possible split pair correlation, is a common metric for this form of reliability.

Then, Cronbach's alpha analysis method in SPSS 2.0 was used in this research to analyse reliability value. Cronbach's alpha is the most popular method of testing for internal consistency in the behavioural sciences. Coefficient alpha is useful for estimating reliability for item-specific variance in a unidimensional test and also for the existence of a single factor or construct has been determined (Cortina, 1993) like this research that measures HOTS skill dimension. Subsequently, the generally accepted minimum standard of reliabilities value is .65 (Ebel & Frisbie, 1991). Thus, the reliability of the test items can be claimed as reliable if the value of reliability is more than .65.

Results of Research

Classification of science tests in reasoning items of TIMSS into three levels of HOTS

For each item in Table 3, the correlation between the students' score who answer correctly and incorrectly and aggregate score on the set for the same domain was used as an index of discrimination (this index will be the usual point-biserial correlation index of item discrimination). If the category is the correct answer, the point-biserial index of discrimination should be higher than .25. The point-biserial index of discrimination for each item on analyzing, evaluating and creating is higher than .25 and significant at the .01 levels. So, it means the category was correct.

Table 3. HOTS's level result of point-biserial correlation.

Items	Point-biserial index	Validity
Analysing		
1b	.41	Valid
2	.81	Valid
5b	.48	Valid
7	.81	Valid
12a	.82	Valid
12b	.43	Valid
Evaluating		
1a	.88	Valid
4	.43	Valid
5a	.35	Valid
8	.83	Valid
10	.86	Valid
11	.74	Valid
Creating		
3	.80	Valid
6	.54	Valid
9a	.80	Valid
9b	.82	Valid

Note: correlation is significant at the .01 level



The degree of difficulty of classified reasoning items of TIMSS

The degree of difficulty scale: < .20 is very difficult; .20 – .50 is good; .50 – .80 is best; and > .8 is very easy (Frank, 1962 in Suruchi & Rana, 2014). The degrees of difficulty in Table 4 of analyzing, evaluating and creating test items exhibited similar trend based on the p-value of each item in point-biserial correlation Table 3, which were in the range .20 to .50 (good) of difficulty index.

Table 4. Result of degree of difficulty, reliability and validity analysis.

Items	Degree of difficulty	Reliability	Validity
Analysing		.69	
1b	.30		.42
2	.25		.82
5b	.27		.48
7	.25		.82
12a	.26		.82
12b	.30		.43
Evaluating		.77	
1a	.32		.74
4	.25		.55
5a	.27		.35
8	.28		.84
10	.30		.87
11	.30		.64
Creating		.73	
3	.21		.82
6	.22		.49
9a	.21		.59
9b	.21		.82

Reliability and validity of classified reasoning items of TIMSS

Reliability analysis showed that the total of Cronbach's alpha value of analyzing, evaluating and creating are more than .65 which is acceptable and indicating a high level of internal consistency (Ebel & Frisbie, 1991). And validity analysis result in Table 4 shows that Pearson correlation value in each item is more than r table value (.13). It means that each test item has significant validity. So, the classified science test items of TIMSS are good to use as assessment tools to measure HOTS of students.

Discussion

In order to examine item discrimination of reasoning items of TIMSS item to be levels of HOTS (analyzing, evaluating, and creating) items, we used point biserial correlation method. The point-biserial correlation is the correlation between the right/wrong scores that students receive on a given item and the total scores that the students receive when summing up their scores across the remaining items. It is a special type of correlation between a dichotomous variable (the multiple-choice item score which is right or wrong, 0 or 1) and a continuous variable (the total score on the test ranging from 0 to the maximum number of multiple-choice items on the test). As in all correlations, point-biserial values range from -1.0 to +1.0. A large positive point-biserial value indicates that students with high scores on the overall test are also getting the item right (which we would expect) and that students with low scores on the overall test are getting the item wrong (which we would also expect). A low point-biserial implies



that students who get the item correct tend to do poorly on the overall test (which would indicate an anomaly) and that students who get the item wrong tend to do well on the test (also an anomaly); (Varma, 2015). Point-biserial correlation can exhibit how much prognostic power an item has and how the item contributes to divinations by conjecturing the correlation between each test item and the total test score. The statistic is helpful for verifying the relative performance of different groups or individuals on the same item (McCowan, 1999).

The result of point-biserial correlation has positive value not only on each HOTS test item of one group but also on all of the group. The result in Table 3 shows that point biserial correlation value in each item has wide enough range that is more than 0.25. The range within .40 to .90 that is all items of HOTS test indicating that students with a high score on each item test are getting the item right too. That means each test item significantly correlate to the total test score in each group, and each item in each group is significantly discriminated. This fact can be strengthened by reason if items with higher point-biserial correlations are more highly discriminating, while those with lower point-biserial correlations are less discriminating (Osterlund, 1998).

The p-value of an item provides the proportion of students that got the item correct and is a proxy for item difficulty (or more precisely, item easiness). The higher the p-value, the easier the item. Problematic items (items with a low point-biserial correlation) may show high p-values, but the high p-values should not be taken as indicative of item quality (Varma, 2015). Difficulty value of an item may be defined as the proportion of the certain sample of subjects who actually know the answer to an item (Freeman, 1962). This p-value (degree of difficulty) means the proportion of student answer correctly in the formula of point-biserial correlation. Thus, the degree of difficulty is one of the components to measure the value of point-biserial correlation. The correlation they have is inversely proportional. Therefore, the value of the degree of difficulty in each item in this research was lower than the value of point-biserial correlation itself as mentioned before theoretically. This can be proved by looking at the result above while the value of point-biserial correlation indicates more than .40 in each item, whereas the degree of difficulty value is not more than .30 in each item. The value of the degree of difficulty based on the result in Table 4 exhibited that degree of difficulty value range is $\geq .21$ and $\leq .32$. This result shows us that each test item has a good value of the degree of difficulty. Thus, it means every item is neither easy nor difficult. In the other word, we can use them to be an instrument of the test in science class.

Beside value of the degree of difficulty, reliability is very important to educational assessment to judge whether or not good a test item. This also can be strengthened by the statement if achievement tests are one of the most important aspects of teaching – learning process and the two most important characteristics of an achievement test are its reliability and content validity (Suruchi & Rana, 2014). This research used reliability and validity test to determine whether HOTS-identified science test item from TIMSS can be used as an assessment tool.

Even though, reliability and validity have no independent correlation in its application. The intercourse they have generally appeared in a way that makes reliability the foregoing need. It is due to an assessment that does not have high reliability and cannot have high validity; if the accuracy of assessment is biased due to the impact of a number of different factors, accordingly the degree to which it measures what it is expected to measure must also be biased. Nevertheless, this reason tends to bring out to attempts to add reliability which mostly means closer and closer specification, and use of methods that have the least error. It affects in collecting and using a limited range of evidence, heading to a decrease in validity. Besides that, if validity is added by lengthening the range of the assessment to put outcomes, afterward reliability is possible to down, because these aspects of achievement are not easily assessed. Yet, while this is like that, for the summative assessment, there has to be a compromise between reliability and validity. Whilst, the data are used for formative assessment, validity is foremost and reliability is less important (Harlen, 2004). Since reliability and validity are not independent of each other – and increasing one tends to decrease the other - it is useful in some contexts to refer to dependability as a combination of the two. The approach to summative assessment by teachers giving the most dependable result would protect construct validity while optimizing reliability (Harlen, 2004).

As Sadler (1989) declared that concern to the validity of decision about individual test should take privilege rather than a concern to the reliability of scoring in any context where the tension is on diagnosis and improvement. Then reliability will follow as a corollary. So, it is better to use the different assessment information for both of them. Thus, we use both reliability and validity test to ensure dependability. The recognition of the interaction between validity and reliability means that, while it is useful to consider each separately, what matters in practice is the way in which they are combined (Harlen, 2004). This has led to the combination of the two in the concept of dependability (William, 1993 in Harlen, 2004). This case can be expressed, as Reliability + Validity = Dependability (James, 1998 in Harlen, 2004). However, there is no simple sum to be calculated here. Since, as noted above, it is not



possible to have high reliability and high validity, it is necessary to consider the balance of priorities. In deciding the relative importance of the two components of dependability, the purpose of the assessment has to be taken into account (Harlen, 2004). Table 4 showed that the reliability and validity value has high value. As a result, this research has high dependability as well. Thus, one can say from this fact that this research is in balance and very good to use as an assessment tool in the science classroom.

Conclusions

Science tests in TIMSS reasoning items can be classified into three levels of higher order thinking skills (HOTS) with value of point-biserial index of discrimination for each item on analyzing, evaluating and creating was higher than 0.25 and significant at the 0.01 level. The classified science test items of TIMSS were good to use as assessment tools to measure HOTS of students due to significantly have high validity value (more than .128), a high level of internal consistency (more than .65) and the difficulty index in good range (0.2 – 0.32). Based on this research results, it is important for teachers in Indonesia to use the classified reasoning item test of TIMSS as assessment tools in measuring higher order thinking skills of students. Furthermore, students should be trained to use such reasoning item test of TIMSS as much as possible to minimize students' error in answering higher order thinking questions. Moreover, this research is also useful for science teachers globally that understanding the taxonomy of questions and good practice strategies for classifying the questions level may help them formulate assessment tools that not only stimulate higher order thinking skills of students but also measure their cognitive level properly.

Acknowledgement

This research was supported by Indonesia Endowment Fund For Education (LPDP). The authors would like to thank Takuya Baba and Ayami Nakaya for some suggestions improving the data.

References

- Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K.A., Mayer, R. E., Pintrich, P. R., Raths, J., & Wittrock, M. C. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's Taxonomy of Educational Objectives* (Complete edition). New York: Longman.
- Barnett, J. E., & Francis, A. L. (2012). Using higher order thinking questionnaire foster critical thinking: A classroom study. *Educational Psychology, 2*, 201-211. doi:10.1080/01443410.2011.638619.
- Cortina, J. M. (1993). What is Coefficient Alpha? An examination of theory and applications. *Journal of Applied Psychology, 78* (1), 98-104.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York, NY: CBS College Publishing.
- Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Englewood Cliffs, NJ: Prentice-Hall, Inc.
- Freeman, F. S. (1962). *Theory and practice of psychological testing*. New Delhi, India: Oxford & Ibh Publishing.
- Harlen, W. (2004). A systematic review of the evidence of reliability and validity of assessment by teachers used for summative purposes. In *Research Evidence in Education Library (issue 4)*, pp. 1-89). London, United Kingdom: EPPI-Centre, Social Science Research Unit, Institute of Education.
- Hotiu, A. (2006). *The relationship between item difficulty and discrimination indices in multiple-choice tests in a physical science course*. Boca Raton, FL: Florida Atlantic University.
- King, F. J., Goodson, L., & Rohani, F. (2015). *Higher order thinking skills*. Tallahassee, FL: Florida State University.
- Krathwohl, D. R. (2002). *A revision of bloom's taxonomy: an overview*. Columbus, OH: The Ohio State University, College of Education.
- Martin, M. O., Mullis, I. V. S., Foy, P., & Stanco, G. M. (2012). *TIMSS 2011 international results in science*. Chesnut Hill, MA: TIMSS & PIRLS International Study Center.
- McCowan, R. J., & McCowan, S. C. (1999). *Item analysis for criterion-referenced tests*. New York, NY: Center for Development of Human Services, Buffalo State College (SUNY).
- Miller Analogies Test. (2012). *MAT reliability and validity*. San Antonio, TX: NCS Pearson, Inc.
- Ministry of Education and Culture. (2013). *Peraturan menteri pendidikan dan kebudayaan republik indonesia nomor 54 tahun 2013* [Regulation of Minister of Education and Culture of Republic of Indonesia No. 54, 2013]. Jakarta, Indonesia: Kementerian Pendidikan dan Kebudayaan.
- Ministry of Education and Culture. (2015). *Peraturan Menteri Pendidikan Dan Kebudayaan Republik Indonesia Nomor 5 Tahun 2015* [Regulation of Minister of Education and Culture of Republic of Indonesia No. 5, 2015]. Jakarta, Indonesia: Kementerian Pendidikan dan Kebudayaan.
- Mullis, I. V. S., & Martin, M. O (Eds.). (2013). *TIMSS 2015 Assessment Frameworks*. Chesnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA).



- Organisation for Economic Cooperation Development. (2015). *Education in Indonesia: Raising to the Challenge*. Paris, French: OECD Publishing.
- Osterlund, S. J. (1998). *Constructing test items: multiple-choice, constructed-response, performance, and other formats* (2nd Ed.). Boston, MA: Kluwer Academic Publishers.
- Pusat Kurikulum dan Perbukuan. (2016). *Laporan Hasil Ujian Nasional 2015* [Report of national exam result]. Jakarta, Indonesia: Kementerian Pendidikan dan Kebudayaan.
- Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18, 119-144.
- Suruchi & Rana, S. S. (2014). Test item analysis and relationship between difficulty level and discrimination index of test items in an achievement test in biology. *Indian Journal of Research*, 3(6), 56-58.
- Van den Berg, G. (2008). The use of assessment in the development of higher order thinking skills. *Africa Education Review*, 1(2), 279-294.
- Varma, S. (2015). *Preliminary Item Statistics Using Point-Biserial Correlation and p-Value*. Morgan Hill, CA: Educational Data Systems.
- Wasis. (2014). Analyzing physics items of UN, TIMSS, and PISA-based on higher-order thinking and scientific literacy. In Sutrisno, H., Dwandaru, W.S.B., Krisnawan, K.P., Darmawan, D., Priyambodo, E., Yulianty, & E. Nurohmah, S. (Eds.), *Proceeding of International Conference on Research, Implementation and Education of Mathematics and Sciences 2014* (pp. 147-154). Yogyakarta, Indonesia: Yogyakarta State University.

Received: October 05, 2017

Accepted: February 15, 2018

Anjar Putro Utomo

M.Ed., Lecturer, Study Program of Science Education, University of Jember, 68121 Indonesia.
E-mail: anjar_pu.fkip@unej.ac.id

Erlia Narulita

Ph.D., Assistant Professor, University of Jember, Indonesia.
E-mail: erlia.fkip@unej.ac.id

Kinya Shimizu

Ph.D., Professor, Graduate School of IDEC, Hiroshima University, 7398511 Japan.
E-mail: kinyas@hiroshima-u.ac.jp

