Research and Evaluation in Education
Volume 2, Number 1, June 2016 (pages 92-107)
Available online at: http://journal.uny.ac.id/index.php/reid

DEVELOPING AN ASSESSMENT INSTRUMENT OF JUNIOR HIGH SCHOOL STUDENTS' HIGHER ORDER THINKING SKILLS IN MATHEMATICS

[1]Samritin; [2]Suryanto
[1]Muhammadiyah University of Buton; [2]Yogyakarta State University
[1]samritin55@yahoo.co.id; [2]suryauny@yahoo.com

Abstract

This study is a research and development study. It aims to produce an instrument for assessing junior high school (JHS) students' higher order thinking skills (HOTS) in mathematics. Its procedure consists of nine steps: (1) Constructing the test specification; (2) writing test items; (3) analyzing test items; (4) conducting the first tryout; (5) analyzing the results of the first try out; (6) revising the test; (7) assembling the test; (8) conducting the second tryout; and (9) analyzing the results of the second tryout. The instrument content validity was obtained through the focus group discussion (FGD) forum, and Delphi technique. The construct validity was found out through the tryout data analysis. The instrument tryout was conducted twice involving 264 participants in the first tryout and 821 participants in the second tryout. The results of the study indicate that the instrument for assessing JHS students' HOTS in mathematics has met the validity and reliability criteria. From the results of the content validity analysis, it can be concluded that the instrument is valid, and it was supported by the items validity indices above 0.79. From the results of the construct validity analysis, it can be concluded that the instrument is valid, as indicated by the value of $\chi^2 = 67.69$, with p-value = 0.10, Root Mean Square Error of Approximation (RMSEA) = 0.03, supported by Goodness of Fit Index (GFI) of 0.97, Normed Fit Index (NFI) of 0.95, and Adjusted Goodness of Fit Index (AGFI) of 0.95. The instrument reliability is 0.88. The developed instrument for assessing HOTS in mathematics consists of 12 items, each of which is of essay test type. The test items have difficulty indices in a range of $0.30 \leq P_i \leq 0.7$.

Keywords: *assessment instrument, higher order thinking, junior high school, mathematics*

How to cite item:
Samritin, S., & Suryanto, S. (2016). Developing an assessment instrument of junior high school students' higher order thinking skills in mathematics. *Research and Evaluation in Education, 2*(1), 92-107.
doi:http://dx.doi.org/10.21831/reid.v2i1.8268

Introduction

The development of thinking skills is an important aspect in education. Byrnes (2008, p.42) states that in Vygotsky's view, thinking skills develop from the lowest level to a higher level. Therefore, school is expected to facilitate the development of students' thinking skills of the lower level to higher level.

Higher level thinking skills or higher order thinking skills (HOTS) can be defined as a cognitive process that involves analysis, synthesis, and evaluation (Stanley & Moore, 2010, p.10). A student who develops his HOTS will have analytical acuity, the ability to synthesize, and good evaluation capabilities. HOTS in mathematics can be defined as the ability to perform mathematical processes or complex tasks or math problems involving connection, problem solving, and mathematical reasoning.

Connection is the ability to see and create linkages among mathematical ideas, between mathematics and other subjects, and between mathematics and everyday life (Kaur & Lam, 2012, p.2). Further, de Lange (1999, p.15), Atkin (2003, p.15), and Shafer and Foster (1997, p.1) classify connection as a second-level math skills. Connection abilities consist of: (1) The ability to make or explain mathematical relationships between concepts or between concepts of mathematics and the real world or between mathematics and other disciplines; and (2) the ability to integrate information and choose different procedures or strategies in solving problems or offering more than one approach to solve a problem.

Solving a problem means finding a way out of a difficulty, a way round of an obstacle, attaining an aim which is not immediately attainable (Polya, 1981, p.ix). To solve a problem means to find such an action (Polya, 1981, p.117). A problem is a situation in which an individual or group is called upon to perform a task for which there is no readily accessible algorithm which determines completely the method of solution (Lester, 1980, p.287). Accordingly, a task is a problem when there is no readily accessible algorithm to reach the solution. In solving a problem, the correct answer could be more than one, and so could the strategies to solve it. The strategy or way to solve a problem can vary but each way produces a correct solution.

The mathematical reasoning involves gathering evidence, making conjectures, establishing generalizations, building arguments, and drawing logical conclusions (Peressini & Webb, 1999, p.156). Formal mathematical reasoning includes reasoning or evidence, which is, a logical conclusion based on assumptions and definitions. The mathematical reasoning often begins with exploration, making allegations, and comes to a conclusion (NCTM, 2000, p.342). Reasoning is an important aspect in mathematics (NCTM, 2009, p.402). Mathematical reasoning is essentially about development, justification, and use of mathematical generalization (Russel, 1999, p.1). Creating generalizations also enables problem solving, as generalizations support learners to see the underlying structure of the problem and the bigger class of problems or ideas that it instantiates. Therefore, mathematics teaching and assessment need to consider the development of students' mathematical reasoning.

Complex problem solving and mathematical reasoning is classified as a third-level math skills (highest level skills) by Atkin (2003, p.15). The achievement of this level (problem solving and mathematical reasoning) is seen from the students' ability to do mathematization, analyze, justify, communicate, interprete, develop own models and strategies, and make arguments and generalizations.

The afore-mentioned description shows that the development of HOTS in mathematics can be facilitated through the offered stimulus such as math problems that require students to analyze, reason, interpret, present ideas, and find and apply mathematical concepts. Giving a variety of new problems will lead students to explore and synthesize concepts logically as creative steps that can lead them to find the right solution. Of course, the problem or the question must be appropriate with the developmental level of students.

The reform movement in mathematics education puts the emphasis on teaching for understanding the learning and assessment of HOTS (Thomas, Okten, & Buis, 2002, p.1).

This opinion emphasizes changes in mathematics teaching practices in order to facilitate the development of HOTS. This opinion also emphasizes that the students' HOTS should also be considered in the assessment. The results of the assessment will have an impact on the implementation of the teaching and learning process.

Brookhart (2010, p. 9 & 12) states that the results of assessing of HOTS increases students' motivation and achievement. This statement shows the importance of assessing of HOTS. The results of a preliminary study conducted in 18 junior high schools (JHSs) in the Province of South East Sulawesi in January - February 2012 found that HOTS in math has not been assessed. It is seen from the payload skills required in these test items, which are used in school. This preliminary study results in the fact that the test items that require students HOTS are not found in JHS' math tests.

These findings indicate that the assessment instrument that is used in the classroom could not assess the students' skills to a higher level, so that the students' skills at the higher level are not known. This indicates that the test results have not provided sufficient or maximum information about the students' skills. The implication is that the teaching process improvement based on the results of the test is also not optimal.

The description indicates the importance of improving the quality of assessment systems. A good assessment system can provide good information to improve the teaching process. The assessment system is quite good if done in accordance with the appropriate procedures/mechanisms, one of which is the use of appropriate instrument.

A test as an instrument used to obtain the information about students' competence development should have a good quality and be developed in accordance with the procedures of instrument development. The preliminary study indicates that the assessment in the classroom is still not well planned. The assessment instrument used in schools has not been well designed. The tests which are used to assess students' learning outcomes are made without regarding the preparation of the tests. These tests are compiled without the test grating. Classroom assessment found that the emphasis on the results of thinking is more dominating than the thinking proccess of students. The test used takes the form of the multiple-choice items more than the other forms. In the field, it was found that 33% of schools as the subject of the preliminary study used multiple-choice test items only, 16.67% schools used the essay test items only, and 66.67% schools used the essay test, with a percentage of 20% at most.

The multiple choice test is the most powerful tool to measure students' mastery of the subject matter. This form can also be used to measure the competencies of students to a higher level. However, the use of multiple choice test items emphasizes only the results, while the students' thinking processes cannot be known. In addition, it is also not known whether the students' response is a result of their thinking or the result of guessing. The use of multiple-choice tests also resulted in non-habit the student to provide a description of answer or argument in solving the problem. For these reasons, the variation of the test type is needed.

In addition to variations of tests types, the quality of test items should be considered in the assessment. To determine the quality of the test, the test item analysis is required. The analysis of the test item is one important aspect in the implementation of the assessment. The results of the test item analysis provides information about the quality of the tests used. If the test items do not have a good parameters, they cannot provide good information as desired. The preliminary study also found that all schools studied had daily test document containing the analysis of the data on students' mastery learning, but all schools do not have the document on test item analysis. This means that these schools have evidence of the development of students but they are not supported by a quality assessment tool known as parameters such as test items.

The fact does not show the real condition of assessment throughout Indonesian schools, but it does show that there are many problems associated with the implementation

of assessment in the classroom which need to be resolved. The development of assessment instruments of students' HOTS in mathematics becomes an important issue to discuss. This is considering that the tests which are able to assess students' HOTS in mathematics and which have evidence of validity and whose item parameters are reliable and good have not been developed. To resolve these problems, a systematic development research is required.

The result of this development study is an instrument to assess students' HOTS in mathematics. It will have implications for improving the quality of teaching in the classroom. The HOTS assessment result can be used to plan the next teaching and learning which can facilitate the development of students' competence to a high level. The result of this assessment will also provide a positive implication for students to study harder to be able to resolve new challenging problems. The students' habits of resolving the new challenging problems can improve their HOTS.

Based on the background mentioned earlier, the research problem can be formulated as 'What is the result of the development of assessment instruments of junior high school students' HOTS in mathematics like?' In line with the formulated problem, the purpose of this research is to produce an assessment instrument of junior high school students' HOTS in mathematics. The results of this study are expected to: (1) Provide benefits to add insight into the theory of HOTS and the development of assessment instruments of HOTS in mathematics, (2) increase the standard instruments that can be used by teachers to assess students' skills in junior high school mathematics, and (3) be a reference for researchers to conduct similar studies or extend research.

Method

Type of Research

This was a development study, which was aimed at producing an instrument for assessing junior high scool students' HOTS in mathematics.

Procedure of Development

The development procedure used in this study referred to the instrument development procedure proposed by Mardapi (2008, p.88) consisting of nine steps: (1) Developing test specifications, (2) writing test items, (3) reviewing the test items, (4) doing the test tryout, (5) analyzing the test, (6) improving the test, (7) assembling the test, (8) carrying out the test, and (9) interpreting the test results. In this study, the step of the development was divided into two stages: Design phase and test tryout phase. The design phase included activities in the first step to the third step and the test tryout phase included activities of the fourth step to the ninth step. In the tryout phase, the fourth step was called the first tryout and followed by analysis while the eighth step was called the second tryout and followed by analysis.

The Design Phase

At this stage, the activities carried out were (1) developing a test specification, (2) writing test items, and (3) examining and repair the test items.

Developing Test Specification

The test specification contained a description of the overall characteristics of the test. This step included activities of (a) specifying the purpose of the test, (b) designing the test blue print, and (c) selecting the test form. The test specification served as a practical manual for test developers to plan the content of the subjects tested, the aspects of behavior to be measured, the test form, and test length.

The test blue print was presented in the form of a matrix that contained the components which consisted of: The material which was tested, measurable aspects of behavior, and cognitive levels to be measured. The cognitive aspect to be measured in this study was determined based on the operational definition of HOTS in mathematics. The cognitive aspect consisted of (1) connection (L2) and (2) problem solving and mathematical reasoning (L3). The aspects of content or material were determined based on the study of mathematics that supported the

achievement of competency standards (CS) and basic competence (BC) in the content standards of mathematics education for grade eight students of junior high school. The material tested, the cognitive behavior, and the cognitive aspect to be measured are outlined in Table 1. The test developed in this study was an esay test.

Test Items Writing

The number of the test items made for each indicator was at least one item, which was tailored to the cognitive aspects measured. The writing of the test items also considered their compliance with the HOT test criteria, namely using new materials (novelty), as Brookhart (2010, p.25), writes that a test item requires a complex thought, using simple sentences but clearly targets the question and uses the good and grammatical Indonesian. Each item was accompanied by an answer key and scoring guidelines or rubrics. The result at this stage was called the initial draft or draft-1.

Reviewing and Improving Test Items

The test items that had been written (Draft-1) were reviewed by experts through a focus group discussion (FGD). The experts were four mathematics education experts and four experts in the field of measurement. This activity was intended to obtain content validity and was conducted on July 26, 2012 at the Graduate School of Yogyakarta State University, Indonesia.

The review activity included reviewing the test blue print, answer key, and scoring guidelines. In general, the review of the test items consisted of test content, test item construction, and language aspects. The judgement and input from the experts in both oral and written forms were subsequently analyzed. Based on the results of the review, Draft-2 was obtained. Draft-2 was reviewed by experts by using the Delpi techniques. The review involved five mathematics education experts. At this stage, a valid test was obtained and it was called Draft-3. Draft-3 was then assessed quantitatively by six experts. The results of this quantitative assessment were analyzed and a validity index was obtained. After assessed by experts, the test items were assembled into a test package. In the test package assembling, the test items were then arranged from easy to difficult items. It is intended to reduce the anxiety of tryout participants. The tests that had been assembled were subsequently tested to obtain the characteristics of the empirical tests.

Tryouts Phase

Design, Subjects, and Tryout Schedule

The tryout activity of the products consisted of two phases: The first tryout and the second tryout. These activities were carried out in the province of South East Sulawesi. The subjects in this study were class VIII students of junior high school. They were selected based on the needs of the development.

Table 1. Blue print

Materials	BC	Indicator	Level
Operations on Algebra	1.1	Solving problems using operations on algebra	L2
Relations and Functions	1.3	Solving problems associated with the relations and function.	L2, L3
Function value	1.4	Solving problems related to the value of the function.	L2, L3
Straight Line Equation	1.6	Solving problems related to the gradient, equations, and graphs straight line	L2
Linear Equation Systems with Two Variables	2.2	Assessing mathematical models of the problems associated with LESTV.	L3
		Making a mathematical model of the problems associated with LESTV.	L3
	2.3	Interpreting LESTV into real-world situations.	L3
		Solve problems related to the LESTV.	L3
		Assessing the settlement of LESTV truth.	L3

They were International-Standard Pioneering School or *Rintisan Sekolah Bertaraf Internasional* (RSBI)/ex-RSBI and also National-Standard School or *Sekolah Standar Nasional* (SSN) students. The number of the schools used was adjusted with the number of test tryout subjects required to take the tests. Crocker and Algina (1986, p.322) state that the acceptable number of participants in the is 200. Moreover, Muraki and Bock (1998, p.35) explain that a good minimum limit for restricted testing activities is as many as 250 people, and for general purposes, a minimum of 500 people are required.

The tryout was conducted twice and it was preceded by a readability test, which was involving 10 year eight students of junior high school and a math teacher of Junior High School (JHS) 6 Raha. The first tryout was conducted in the regency of Baubau, Southeast Sulawesi. This activity involved 264 year eight students of State JHS 1 and State JHS 2 Baubau, and was conducted on November 23 to 25, 2013. The first tryout result was analyzed and used as the basis to revise the instrument.

The test which had been revised based on the first tryout data analysis was tested on a large scale. It was the second tryout, which was held on 2nd to 10th December 2013. This activity involved 821 eight grade students from four schools, namely State JHS 2 Baubau, State JHS 1, State 2 JHS 2, and State 3 JHS 3 Raha. The analysis of the data from the second tryout was intended to see if the test had satisfied the specified criteria or not. The results of the second tryout data analysis showed that the test satisfied the specified criteria. Therefore, the test was not revised and became the final draft.

Data Type, Instruments and Data Collection Techniques

The data in this study were quantitative, in the form of scores which were given to students' responses to the tried out test. In order to obtain the data, instrument was used. The instrument was obtained through the process of judgement and it was revised based on the suggestions from experts.

Data Analysis Techniques

The qualitative data obtained from the experts were analyzed to answer the question of 'whether the product was valid or not'. Instrument validation results based on the judgement of experts were analyzed qualitatively and were revised if necessary.

The quantitative data obtained from the experts' assessment of the test items were analyzed using formula Aiken validity indices (Aiken, 1985, p.132) and the results were used as an evidence of content validity quantitatively. Aiken (1985, p.134) sets the lowest value of validity index depending on the number of experts and the criteria used. The lowest value for six experts and five criteria is 0.79.

The quantitative data on the results of test tryouts were used to answer the question of 'whether the test satisfied the criteria of construct validity, reliability, and item parameters'. The validity of the test which was based on empirical data was analyzed using confirmatory factor analysis (Mueller, 1996, p.112). The instrument was considered valid if the model fits the data. The criteria which were used to make decisions that the model fit to the data were based on: (1) p-value of Chi-squre (X2)> 0.05, (2) the Root Mean Square Error of Approximation (RMSEA) < 0.5 (Schumacker & Lomax, 2004, p.82).

The instrument reliability coefficient criterion used was minimum 0.7 (Nunnally, 1981, p.245; Urbina, 2004, p.137). The assessment of the instrument tryout results was conducted by two assessors. The calculation of the inter-rater consistency used Cohen's Kappa formula and the calculation of the instrument reliability based on the verified data used the alpha Cronbach's formula.

The item parameter of the instruments according to CTT was seen from the difficulty and discrimination indices. However, the discrimination index of the criterion-referenced test does not affect the quality of the test, so that the item parameter in this study was only the difficulty index. The criteria used to determine the item difficulty index was 0.3 to 0.7 (Allen & Yen, 1979, p.121). The estimation of the test item difficulty (Pi) referred to the formula by Nitko and Brookhart (2007, p.324), as follow:

$$P_I = \frac{score\,average\,of\,item - minimum\,score\,of\,item}{maximum\,score\,of\,item - minimum\,score\,of\,item}$$

Findings and Discussion

The indicators derived from the selected Basic Competenties were loaded in instrument blue print, which is a reference in writing instrument items. The instrument consists of items that have been written, called Draft-1. Draft-1 consists of 18 items, each of which is in the form of an esay test.

Instrument Validation Results

The initial draft of the instrument or Draft-1 which had been developed, subsequently handed over to the experts to be reviewed. The review of the test was conducted through focus group discussion (FGD) forum. The program involved eight experts consisting of four experts of mathematics education and four experts of educational measurement. In this activity, the experts did the review of the draft of the developed instruments including scoring the rubric and blue print of the instrument. In general, the experts' suggestions were: (1) A few items of the instrument must be replaced or revised because the problems are not in accordance with the criteria of HOT test in mathematics; and (2) a few items of the instrument need to be revised to make them more suited to the conditions of students.

Based on the experts suggestions, the instrument was revised. The results of this revision was then discussed again by experts through Delpi techniques. The discussion was was conducted several times with each expert to obtain a valid test. The discussion at this stage resulted in Draft-3 of intrument. Draft-3 consisted of 14 items that had been declared valid by the experts.

Draft-3 was produced through Delpi techniques and was assessed by six experts. The scores given were based on the experts' point of view of the relevance of the indicator. There were five criteria used in the assessment: The score of 1 if the item was not relevant, the score of 2 if the item was not relevant, the score of 3 if the item was less relevant but could be used, the score of 4 if the item was relevant, and the score of 5 if the item was very relevant. Based on the results of the quantitative judgement, further validity index of each item was calculated using the formula of Aiken (1985, p.132). Steps were taken to obtain the validity of the Aiken index by first calculating the number of assessors in the *-ith* criterion of each item, and then index V Aiken was calculated. The results of the calculation are presented in Table 2.

Table 2. The number of assessors on each criterion and V Aiken indices for each item

Item	Criteria					Sum of Experts	V
	1	2	3	4	5		
1	-	-	-	3	3	6	0.875
2a	-	-	-	4	2	6	0.833
2b	-	-	-	4	2	6	0.833
3a	-	-	-	4	2	6	0.833
3b	-	-	-	2	4	6	0.917
4	-	-	-	2	4	6	0.917
5	-	-	-	3	3	6	0.875
6	-	-	-	4	2	6	0.833
7	-	-	-	3	3	6	0.875
8	-	-	-	4	2	6	0.833
9	-	-	-	2	4	6	0.917
10	-	-	-	1	5	6	0.958
11	-	-	-	2	4	6	0.917
12	-	-	-	2	4	6	0.917

Table 2 shows that the validity of each item index is above the specified minimum criterion, 0.79 (Aiken, 1985, p.134). The criteria were established by Aiken to six experts with the scale of 5. Based on the index of every item, it was concluded that the instrument was valid. Therefore, it was determined that the instrument was ready to be tried-out.

Product Tryout Results

The First Tryout Results

The results of the first tryout were scored. The scoring was done using a scoring rubric. Each item had a scoring rubric in accordance with that item. In developing the scoring rubric, the fairness aspect was considered so that students were not disadvantageous in the scoring. The scoring rubric used was the analytic form. Before used, the scoring rubrics were reviewed by the experts through focus group discussions and Delpi techniques.

The scoring was performed by two assessors. The use of two assessors in scoring was intended to avoid the effect misinterpretation of the students' answers, the effect of fatigue, and other effects. The scoring was done on the students' answers by scoring one item at a time. It resulted in two data scores. Therefore, to produce a single data score, two assessors verified the data.

The verification of scores was performed on the different scores by the assessors. The verification of scores was performed by the assessors by reviewing the answer sheets together. Based on the results of the verification an accurate data score was obtained. The verified data score was then used to analyze the item parameter, reliability, and validity based on empirical data.

Items Test Parameter on the First Tryout

The difficulty and discrimination indices are two item test atributes on the CTT analysis. However, in the criterion-related test, the discrimination index is not considered in the selection of items. Thus, in this study, the parameter analyzed was only the item difficulty index (Pi). The results of the analysis of the item difficulty of the instrument are presented in Table 3.

Table 3. Test item difficulty indices on the first tryout

No	Item	Difficulty (P_i)
1	1	0.34
2	2a	0.55
3	2b	0.59
4	3a	0.07*
5	3b	0.06*
6	4	0.32
7	5	0.31
8	6	0.30
9	7	0.45
10	8	0.40
11	9	0.43
12	10	0.35
13	11	0.33
14	12	0.35

Table 3 shows that there are two items that indicate the instrument is too difficult to resolve by students. It is seen from the difficulty indices of the items, each of which is less than 0.3, 3a has an index of 0.07 and item 3b has the difficulty index of 0.06. The other test items have indices of difficulty the range of 0.30 to 0.7.

The results of the search to the items that had difficulty indices of less than 0.3 showed that Item 3a was answered by only 51 or 19.3% of the tryout participants. For Item 4a, the number of participants who obtained a score of 1 was as many as 48 students, score of 2 as many as three students, and none of the participants was able to achieve a maximum score of 3. Item 3b was answered by only 46 or 17.4% of the tryout participants. For Item 4a, the number of participants who obtained the score of 1 was as many as 42 students, score of 2 as many as four students, and none of the participants was able to achieve the maximum score of 3. Therefore, Items 3a and 3b were then removed from the test package and not included in the next analysis.

Instrument Reliability in the First Tryout

The reliability of an instrument is related to the measurement error. The scoring which was conducted by more than one rater on the same instrument will provide high reliability if the consistency of scoring is high. This means that an inter-rater measurement error illustrates the magnitude of the inconsistency scores given by the two scorers. The reliability of the instrument in this study was considered from the coefficient of inter-rater Kappa (measure of agreement Kappa) of Cohen and Cronbach's alpha. Technically, the reliability coffiecient estimation was performed with the help of SPSS. The results of inter-rater agreement calculations are presented in Table 4.

Table 4 shows that the consistency of measurements by the two scorers is very high. This is seen from Kappa coefficient at least 0.950 on number 7. The reliability of the instrument in the first tryout of this study was estimated based on data verified coefficient $\alpha = 0.87$. This coefficient is higher than the required minimum reliability coefficient of 0.7 for a good instrument (Nunnally, 1981, p.245; Urbina, 2004, p.137). Thus, based on the results of the first tryout it is concluded that the instrument is reliable.

Table 4. Kappa coefficients of each item based on the results of the first tryout

No.	Item	Kappa Coefficient
1	1	0.983
2	2a	0.976
3	2b	0.971
4	4	0.978
5	5	0.961
6	6	0.950
7	7	0.961
8	8	0.973
9	9	0.988
10	10	0.963
11	11	0.961
12	12	0.967

Analysis of Instruments Validity based on the Data of the First Tryout Results

The construct validity of the instrument was analyzed using factor analysis. Factor analysis used in this study is confirmatory factor analysis (CFA). CFA was conducted by using Lisrel. The data used in the CFA are the verified data. CFA was conducted after the result of the normality assumption testing was obtained. The normality testing results indicate that the data have univariate non-normalities, shown by the p-value of Skewness and Kurtosis of each variable (items instrument). All of the test items have the p-value = 0.00 <0.05. The results of tests of multivariate normality also showed the p-value = 0.00 for Skewness and Kurtosis. This indicates that the data have a multivariate non-normalities. To perform the factor analysis of the data which do not meet the requirements of normality Lisrel, additional data in the form of asymptotic covariance matrix (ACM) were required in order to obtain unbiased estimation results (Schumacker & Lomax, 2004, p.34).

The validity analysis based on empirical data from the first tryout was conducted without including items 4a and 4b because these items have been removed based on the analysis of item parameter according to CTT. The results of the validity analysis showed X2= 65.51 with the p-value = 0.14, and RMSEA = 0.04, which indicates that the model fits to the data. The model is declared fit to the data means the instrument is valid based on empirical data of the first tryout results. Figure 1 and Figure 2 show the fulfillment of these criteria. The results of the analysis are shown in Figure 2, which also shows that the items of the instrument have a significant relationship with the HOTS.

Figure 1 shows that the lowest loading factor of the instrument items on the first tryout is 0.46 and the highest is 0.7. Figure 2 also shows that the loading factor of each item at $\alpha = 0.05$ is significant. This is indicated by the item minimum t-value of 5.72 more than of $t\alpha=0.05 = 1.96$. So, the correlation of the instrument items to HOTS is significant.

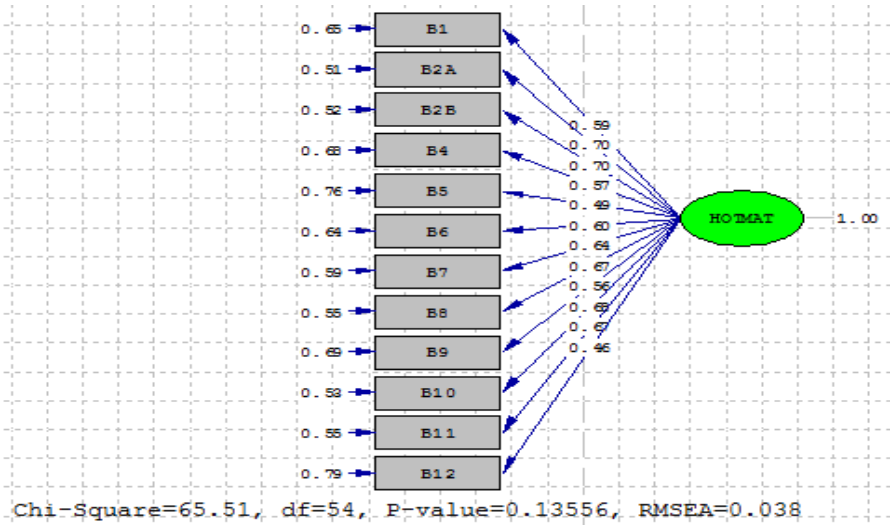

Figure 1. Loading factor the Instrument Items Based on the First Tryout Data

Results of the Second Tryout

Analysis of the Instrument Item Parameter on the Second Tryout

The results of the analysis of the difficulty parameter test items based on data from the second tryout are clearly presented in Table 5.

Table 5. Item difficulty indices based on the second tryout

No	Items	Difficulty (P_i)
1	1	0.35
2	2a	0.59
3	2b	0.61
4	3	0.32
5	4	0.33
6	5	0.36
7	6	0.49
8	7	0.41
9	8	0.48
10	9	0.37
11	10	0.33
12	11	0.39

Table 5 shows that the difficulty para-meters on all test items are in the range of 0.30 ≤ Pi ≤ 0.7 which means that all items have a good parameter. The easiest instru-ment items have the difficulty indices of 0.59 and 0.60. Those items are Items 2a and 2b. Both of these items were formulated from the indicators derived from Basic Competency (BC) 1.3, that is *to understand relationships and functions*. In teaching and learning processes, the basic competency is developed through learning the subject of *relations and functions*. This means that both items were formulated to measure the HOTS of junior high school students on the material *Relations and Functions*.

The most difficult instrument items developed in this study have difficulty indices of 0.32 and 0.33. The item that has the difficulty index of 0.32 is Item3. The item that has the difficulty index of 0.32 is Items 4 and 10. These items are formulated from three indicators derived from two different basic competencies. Item 3 is formulated of the indicators derived from BC 1.3, i.e. *to understand relations and functions*, or to measure JHS students' HOTS on *Relations and Functions*.

Item 4 is formulated of the indicators outlined in BC 1.4, i.e *to determine the value of a function*. To achieve this competence, the students learned the learning material on *Relationships and Function*. Item 4 is an item that is formulated to measure JHS students' HOTS in *Relations and Functions*.

Item10 is formulated from indicators derived from BC 2.3m i.e *to finish mathematical models of the problems associated with the system of linear equations of two variables*. To reach BC 2.3, the students have to learn *Linear Equations System with Two Variables*. Item 10 is used to measure JHS students' HOTS in *Linear Equations System with Two Variables*.

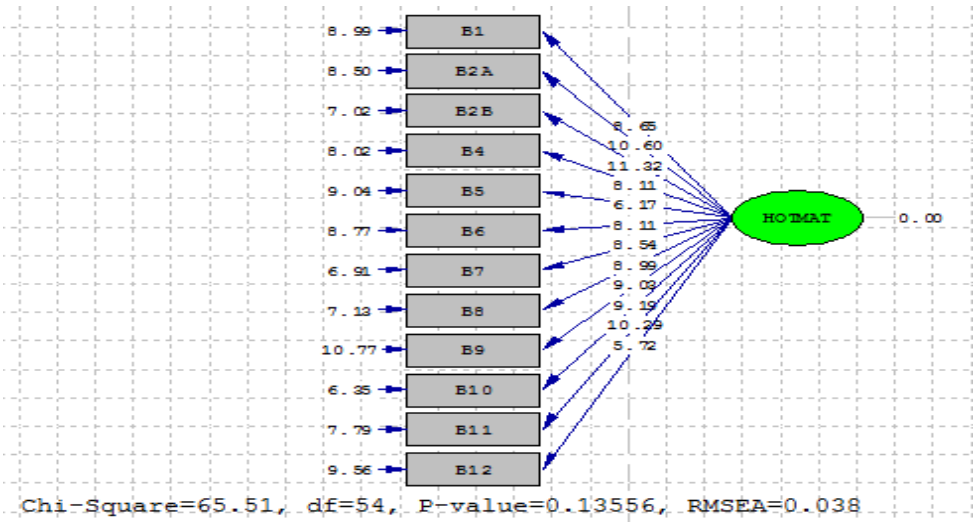

Figure 2. Results of t-value estimation based on the first tryout data

The instrument developed in this study consists of four items on the connection level (L2) and eight items on the problem solving and mathematical reasoning level (L3). The items on L2 are Items 1, 3, 4, and 6. Item 1 is formulated from indicators derived from BC 1.1. Item 3 is formulated from indicators derived from BC 1.3. Item 4 is formulated from indicators derived from BC 1.4. Item 6 is formulated from indicators derived from BC 1.6.

The items on L3 are Items 2a, 2b, 5, 7, 8, 9, 10, and 11. The items on L3 are formulated from the indicators derived from BC 1.3 (i.e. item numbers 2a, and 2b), BC 1.4 (i.e. Item 5), BC 2.2 (i.e. Items 7 and 11), and BC 2.3 (i.e. Items 8, 9, and 10).

Instrument Reliability on the Second Tryout

The reliability of an instrument is related to the measurement error. The scoring done by more than one scorer on the same instrument will provide high reliability if the consistency of scoring is high. The inter-rater consistency in this study was calculated using Cohens' Kappa measure of agreement and the instrument reliability of the verified data was calculated using Cronbach alpha formula. The inter-rater consistency calculation results of the second tryout are presented in Table 6.

Table 6. Kappa coefficients of each item based on the results of the second tryout

No.	Item	Coefficient of Kappa
1	1	0.971
2	2a	0.962
3	2b	0.992
4	3	0.973
5	4	0.980
6	5	0.983
7	6	0.951
8	7	0.982
9	8	0.965
10	9	0.959
11	10	0.949
12	11	0.993

Table 6 shows that the consistency of measurement made by the two scorers is very high. This is evident from its lowest Kappa coefficient of 0.949 on Item 10. The reliability of the instrument based on the verified data on the second tryout, which was calculated by using the formula of Cronbach's alpha, showed a coefficient of 0.88. It is higher than the specified minimum reliability coefficient of 0.7 (Nunnally, 1981, p.245; Urbina, 2004, p.137), so it was concluded that the instrument is reliable.

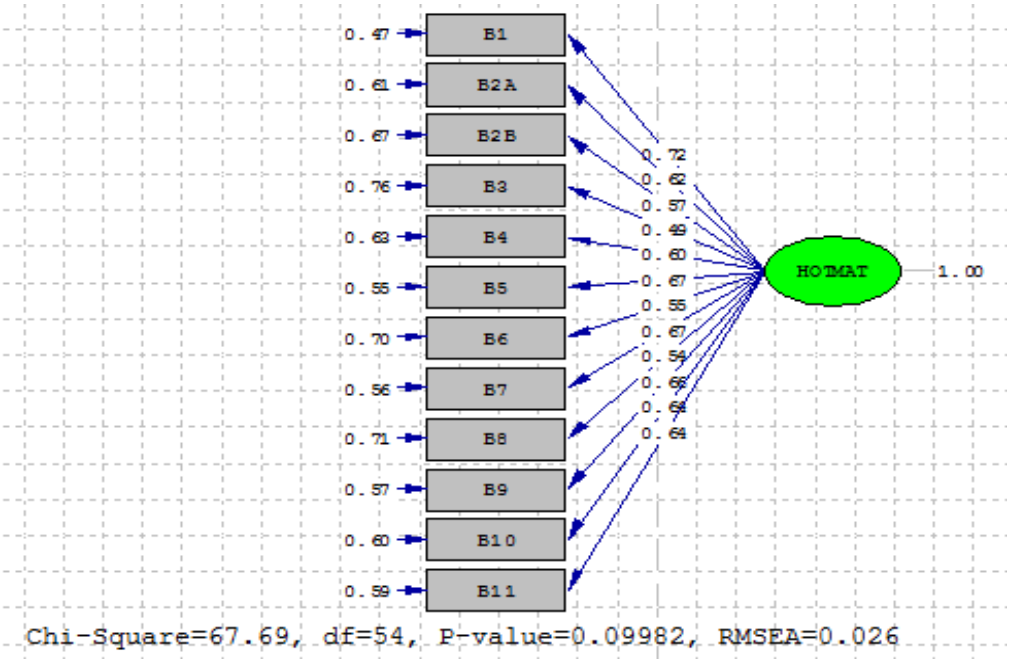

Figure 3. Instrument item loading factor based on the second tryout data

Instrument Validity based on the Results of the Second Tryout

The validity analysis which was based on the empirical data of the second tryout results was conducted using factor analysis. Factor analysis which was used in this study is CFA, which is conducted using Lisrel. The data which were used in the CFA were the verified data. CFA was conducted after the result of normality assumption testing was obtained. The result of the normality assumption testing indicates that the data have univariate non-normalities, as shown by the p-value of Skewness and Kurtosis of each variable. All of the test items have p-value = 0.00 <0.05. The result of the multivariate normality testing also showed p-value = 0.00 for Skewness and Kurtosis. This indicates that the data have multivariate non-normalities. Therefore, it is concluded that the data are not normally distributed.

The confirmatory factor analysis of the second tryout data used additional data which are called asymptotic covariance matrix or ACM as described in the data analysis of the first tryout results. The results of the validity analysis are presented in Figure 3 and Figure 4. Figure 3 shows that the results of the confirmatory factor analysis show X2= 67.69 with p-value = 0.10 and RMSEA = 0.03, which indicates that the model fits the data. The fitness of the model to the data is supported by GFI = 0.97, AGFI = 0.95, and NFI = 0.95, respectively ≥ 0.95 (Schumacker & Lomax, 2004, p.82). That the model was declared fit the data means the instrument is valid based on empirical data.

Figure 3 and 4 show that the items of instrument have a significant relationship with HOTS in mathematics. In the figure, it can be seen that the lowest loading factor of the instrument items in the second tryout is 0.49 and the highest is 0.72. Based on the t-value given in Figure 4, it can be concluded that the loading factor of each item of the instrument is significant at $\alpha = 0.05$ level. The t-value of each items is at least 13.94, more than that of $t\alpha = 0.05 = 1.96$.

Junior High School Students' HOTS in Mathematics

The test scores of tryout results are the source of information about HOTS in mathematics of junior high school students who took the tests in tryout activities. The information about the students' HOTS was obtained through the interpretation of scores. The score interpretation resulted in a value. This value can be presented in the form of numbers or words.

The score interpretation results or assessment results can be used for various purposes such as to improve the quality of learning and to report learning outcomes.

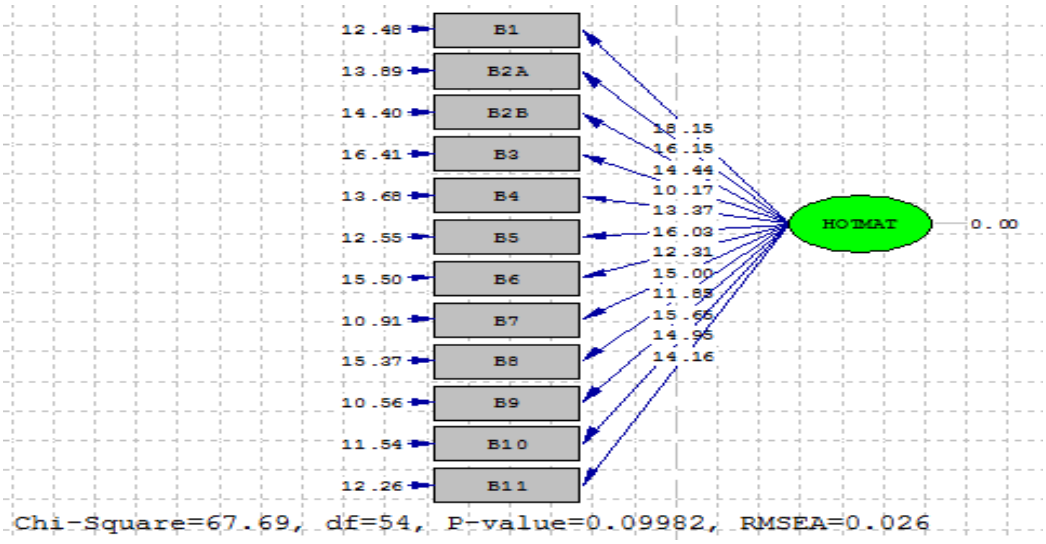

Figure 4. Estimation Results of t-values based on the Second Tryout Data

Reporting students' learning outcomes in each subject in school uses a composite score obtained from several tests. The values obtained from different tests sometimes have different score ranges or scales, so that these values cannot be composited from the raw scores. Therefore, a transformation process of scores from each source into a certain score range or scale is required. Furthermore, these scores may be composited into the final value, in accordance with the desired rules.

The results of assessing the HOTS of junior high school students in mathematics is one of the important sources for composite score reported. When schools use grades 0-10 or 0-100 on the final report, then the scores of the test participants must be transformed into the value of 0-10 or 0-100. This transformation can be performed by using linear transformation by dividing the score of the acquisition with the ideal score, and then the result is multiplied by 10 to obtain a value in the range 0-10 or multiplied by 100 to obtain a value in the range 0-100. In the range 0-10, the highest value obtained by the test participants is 9.39 and the lowest is 0.00. In the range 0-100, the highest score obtained by the test participants is 93.93 and the lowest is 0.00.

The assessment results can also be presented in the form of predicate *very low* to *very high*. Producing a value in the form of a predicate can be done by making a categorization score. The HOTS test in this study has a maximum score of 33 and a minimum score of 0. Thus the range of scores is 36, and the average value is 16.5. The ideal range is divided into six units of standard deviation, resulting in 5.5 as the ideal standard deviation. Based on the ideal average ($\overline{X_i}$) and the ideal standard deviation (Si), the categorization of junior high school students' HOTS in mathematics is: (1) $\overline{X_i}$ + 1.5Si < X or 24.75 < X (very high category); (2) $\overline{X_i}$ < X ≤ $\overline{X_i}$ + 1,5Si or 16.5 < x ≤ 24.75 (high category); (3) $\overline{X_i}$ - 1,5Si < X ≤ $\overline{X_i}$ or 8.25 < X ≤ 16.5 (low category); and (4) X ≤ $\overline{X_i}$ - 1,5Si or X ≤ 8.25 (very low category) (adapted from Azwar, 2009, p.108).

Based on the categorization, it is known that junior high school students' HOTS in mathematics at the second tryout is: (1) Participants who have a very high skill are as many as 9.74%; (2) participants who have a high skill are as many as 25.94%; (3) participants who have a low skill are as many as 29%; and (4) participants who have a very low skill are as many as 35.32%.

The results of assessing junior high school students' HOTS in mathematics show that the dominant value is held by the participants who have low and very low skills, as many as 64.32%. This percentage indicates the number of test takers who have scores no more than the ideal average score. While the test participants who have high and very high ability is only 35.68%. This percentage shows the number of participants who have scores above the ideal score average.

The description shows that the HOTS in mathematics of junior high school students that involved in the second tryout tends to be low. In relation to this, the search of the acquisition results of the scores of the participants was carried out. The result of the search shows that the average score of the students is 13.42. This score is 3.08 lower than the ideal score. The score distribution not normally visible from skewness value is 0.32>0. The slope of the distribution of the scores is shown in Figure 5. Urbina (2004, p.60) argues that the skewness > 0 occurs if most scores are at the low level. This means that the test participants are dominated by those whose score is low.

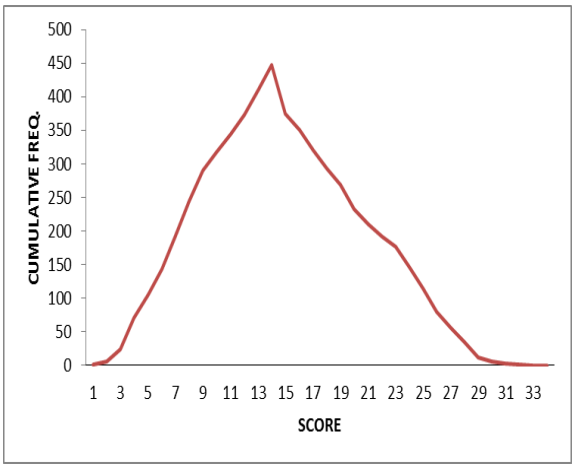

Figure 5. Score distribution curve in the second tryout

The results of the search on the participants' test scores show the highest score of 31 with a frequency (f) = 1 or 0.12%. The lowest score obtained by participants is 0 with the frequency (f) = 1 or 0.12%. The dominant score obtained by the participants is 7 with the frequency of 51 or 6.21%. This means that the number of the test takers who scored 7 is higher than that of those who have other scores. The next dominant score achieved by participants is 6 with the frequency (f) = 50 or 6.09% followed by the score of 3 with the frequency (f) = 47 or 5.72%.

Discussion

The most suitable form of instrument used for assessing students' HOTS in mathematics is essay type test because the students' thinking processes can be determined based on the description of the given answer. An essay type test requires them to demonstrate their knowledge in accordance with the demanded problem. Typically, all forms of tests can be used to assess students' HOTS in mathematics, such as, multiple choice test, but their thinking process cannot be determined. The correct answer chosen by the students in multiple-choice tests cannot reveal whether it is the result of thinking or guessing.

The instrument items for assessing junior high school students' HOTS in mathematics developed in this study has a difficulty index parameter in the range of 0.3 to 0.7. The difficulty index of the items is in a good category. This is due to the development of the instrument which has been through a systematic process and well done. The instrument development process which starts from the preparation of test specifications and then proceed with writing the test items performed by considering various aspects that can affect the students' ability to answer the questions. Those aspects are the suitability of the indicator and test items with the curriculum and students' developmental level, language aspects, and cultural aspects. Another factor affecting the test item parameters developed in this study so that they are in a good category is the arrangement of test items into the test package.

The arrangement of test items into a test package from the simple to the most difficult item is to reduce the anxiety of students in answering the questions. The low anxiety of the students when doing the test allows them to answer according to their ability. The answers which are in accordance with the students' true abilities affect the test item difficulty parameter because the test items are designed in accordance with the level of development and knowledge of students. In the writing of test items, the level of students' knowledge was seen from the content of the curriculum in use.

The instrument which was developed in this study is in the valid category as the implication of a systematic process of instrument development which is done well. The validity evidence consists of content validity and construct validity evidences. The content validity evidence of this instrument is obtained from the experts' judgement and the construct validity evidence is from the analytical results of fitting the model with empirical data. The writing of the test items in this study which is considered representative of curriculum content and fulfillment of the criteria of HOT test in mathematics and the competence of the experts or instrument reviewers are the factors that affect the validity of the instrument. The experts involved in the review and assessment of the instrument items are competent experts in mathematics education and measurement, so that the results of the assessment can be justified. The results of the experts' judgement show that from the content aspect, the instrument is valid, which is visible from Aikens validity index where the lowest is 0.83. The experts have done reviewed and assessed the instrument, so that the resulting instrument really measures what it is supposed to measure, as indicated by fitness of the model to empirical data.

The esay test form must be supported by the scoring guidelines called the scoring rubric. The scoring rubric used in this study was designed as well as possible and validated together with the items of the instrument developed, so that the difference in the scores given by the two scorers was very little, and this is also supported by the lowest Kappa

coefficient of 0.949 of each item. This shows that the inter-rater reliability coefficient is in a very good category. The different scores were verified by two scorers very carefully so as to produce accurate score data with the lowest error of measurement, which is visible from the reliability coefficient of $\alpha = 0.88$. The coefficient is in a good category and has met the criteria. The description shows that the instrument developed in this study has met the reliability criteria viewed from both the inter-rater reliability aspect and the normative aspect.

Junior high school students' HOTS in mathematics tends to be low. This is due to the students' unfamiliarity in working on the problems that require HOTS. The students are accustomed to working on the problems that require low skills so that only some students are capable of achieving the maximum score when solving HOTS problems.

The junior high school students' HOTS in mathematics is mostly low, but some students of State JHS 1 Baubau said that they were happy doing these tests because the problems in the tests were very challenging and arouse their curiosity. The students claimed to agree when the tests or final exams included questions that demanded high level tinking. These students have benefited from this research.

Meanwhile, in the discussions with the mathematics teachers in the tryout, it was found that the teachers agree to use a test that requires HOTS in tests or final exams. It is just that there are difficulties in constructing test items with strict criteria, particularly the novelty criteria of items. They argue that it is very good when there are examples of such test items available. The teachers also claimed that many students, especially those with middle and low ability, will have difficulty to provide the correct answers.

Conclusions and Recommendations

Conclusions

Based on the analysis of the findings, it can be concluded that: (1) The instrument for assessing students' higher order thinking skill in mathematics which is developed in this study consists of 12 items, each of which is of the essay type test item; (2) the test items developed in this study have difficulty indices ranging from 0.3 to 0.7, which means it meets the criteria of a good item parameter; (3) the instrument developed in this study has a reliability coefficient of 0.88, which means that it meets the criteria; (4) the instrument for assessing JHS students' HOTS in mathematics developed in this study is valid, whose evidence is indicated by the item validity index above 0.79, and whose evidence of construct validity based on empirical data is indicated by the value of $\chi 2 = 67.69$, p-value = 0.10, RMSEA = 0.03, GFI = 0.97, NFI = 0.95 and AGFI = 0.95.

Recommendations

Based on the results of the analysis of the findings, it is recommended that: (1) Such an instrument should be developed using standard procedures in order to produce a good HOT assessment instrument; (2) teachers assess the ability of students' thinking to higher levels; (3) mathematics teachers be trained to create HOTS assessment instruments; (4) Department of Education conduct training for teachers in the development of an assessment instrument of HOTS; and (5) other researchers conduct further research in order to increase the number of items of instruments for asessing students' HOTS and for all levels.

References

Aiken, L.R. (1985). Three coefficients to analyzing the reliability and validity of rating. *Educational and Psychological Measurement*, 45, 131-142.

Allen, M.J. & Yen, W.M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.

Atkin, J.M. (2003). *Assessment in support of instruction and learning. Workshop report.* Washington, WA: The National Academies.

Azwar, S. (2009). *Penyusunan skala psikologi (12th ed.)* [Composing psychological scale]. Yogyakarta: Pustaka Pelajar.

Brookhart, S.M. (2010). *How to assess higher order thinking skills in your classroom*. Alexanderia, VA: ASCD.

Byrnes, J.P. (2008). *Cognitive development and learning in instructional contexts*. Boston, MA: Pearson Education.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. New York, NY: CBS College.

de Lange, J. (1999). *Framework for classroom assessment in mathematics*. Retrieved on January 12, 2012 from http://www.fi.uu.nl/catch/products/framework/de_lange_frameworkfinal.pdf

Kaur, B. & Lam, T.T. (Eds.). (2012). *Reasoning, communication and connections in mathematics*. Singapore: World Scientific.

Lester, F.K. (1980). Research on mathematical problem solving. In Shumway, R.J. (Eds.). *Research in Mathematics Education*, pp. 286-323. Reston, VA: NCTM.

Mardapi, D. (2008). *Teknik penyusunan instrumen tes dan nontes* [Test and non-test instruments composing techniques]. Yogyakarta: Mitra Cendekia.

Mueller, R.O. (1996). *Basic analysis of structural equation modeling*. New York, NY: Springer-Verlag New York.

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

National Council of Teachers of Mathematics (NCTM). (2000). *Principles and standards for school mathematics*. Reston, VA: The National Council of Teachers of Mathematics.

National Council of Teachers of Mathematics (NCTM). (2009). *Guiding principles for mathematics curriculum and assessment*. Reston, VA: The National Council of Teachers of Mathematics. Retrieved on 15 January 2013 from http://standards.nctm.org/document/chapter2/content.aspx?id=23273

Nitko, A.J. & Brookhart, S.M. (2007). *Educational assessment of students*. Boston, MA: Pearson Prentice Hall.

Nunnally, J.C. (1981). *Psychometric theory*. New Delhi: McGraw Hill.

Peressini, D. & Webb, N. (1999). Analyzing mathematical reasoning in students' response across multiple performance assessment tasks. In Stiff, L.V. & Curcio, F.R. (Eds.). *Developing Mathematical Reasoning in Grades K-12*. pp. 156–174. Reston, VA: NCTM.

Polya, G. (1981). *Mathematical discovery: On understanding, learning, and teaching problem solving*. New York, NY: John Wiley & Sons.

Russel, S.J. (1999). Mathematical reasoning in the elementary grades. In Stiff, L.V. & Curcio, F.R. (Eds.). *Developing Mathematical Reasoning in Grades K-12*. pp. 1–12. Reston, VA: NCTM.

Schumacker, R.E. & Lomax, R.G. (2004). *A beginner's guide to structural equation modeling (2nd ed.)*. Mahwah, NJ: Lawrence Erlbaum Associates.

Shafer, M.C. & Foster, S. (1997). The changing face of assessment. *Principled Practice in Mathematics & Science Education,* 1(2), 1-12.

Stanley, T. & Moore, B. (2010). *Critical thinking and formative assessment*. Lachmont, NY: Eye On Education.

Thomas, D.A., Okten, G., & Buis, P. (2002). *On-line assessment of higher-order thinking: A java-based extension to closed-form testing*. ICOTS6, 1-4. Retrieved on June 6, 2013 from https://www.stat.auckland.ac.nz/~iase/publications/1/6d4_thom.pdf

Urbina, S. (2004). *Essential of psychological testing*. Hoboken, NJ: John Wiley & Sons.