TES Construction Analysis and Student Learning Results “Masyarakat Damai” Junior High School Gunungsitoli
Amin Otoni Harefa1
1Lecture, IKIP Gunung Sitoli, North Sumatra Province – Indonesia

Abstract: The test construction analysis is the study of a set of learning outcomes, whether a set of test results is feasible as a good test in a valid, reliable, objective, and practical sense. Thus a good test can measure student learning outcomes appropriately, correct weaknesses and shortcomings, improve the effectiveness of the teaching-learning process, support the fluency and success of the teaching-learning process, and can assist school responsibilities in expediting curriculum implementation, assisting schools in improve the quality of graduates, increase the credibility of schools with the presence of teachers who have skills in preparing the problem.

Keywords: test construction, learning outcomes.

INTRODUCTION

Education is a dynamic thing; therefore education needs to get attention both in business development and quality improvement in accordance with the development of science and technology. This is in accordance with Law No.20 of [11] on National Education System, Chapter II Article 3 which reads: National Education function to develop the ability and form the character and civilization of a dignified nation in order to educate the life of the nation that aims to develop human beings who believe and piety to God Almighty, noble, healthy, knowledgeable, capable, independent and become a democratic and responsible citizen.

To achieve the National Education system, teachers are required to stimulate students' learning motivation and be able to carry out their duties as educators, teachers, and trainers well, and able to think of appropriate strategies, appropriate and precise so what is the expected learning objectives can be achieved as much as possible. To achieve the maximum objective, one of them as proposed by Majid [1] ”lies in the competency component of proper learning management which includes: (1) preparation of learning planning; (2) implementation of teaching and learning interaction; (3) assessment of learning achievement of learners; (4) implementation of follow-up assessment results “.

To gauge the success of the implementation of the evaluation of learning outcomes should be applied properly, precisely and correctly. According Muri [2] ... evaluation of education can be: 1. Evaluation context / goals / policies. 2. Evaluation of inputs, such as evaluation of learners, educators, infrastructure and facilities, curriculum / program, and environmental input. 3. Evaluation of the process, ie the evaluation of the ongoing process or educational or learning activities. 4. Evaluate results / products. 5. Evaluate "outcomes" (impact).

Implementation of a good evaluation, appropriate and true closely related to the curriculum and its implementation adjusted to the demands of the current curriculum. Thus, as well as carrying out various evaluations. The curriculum improvement from year to year, especially the curriculum changes from curriculum oriented to the content of learning to a competency-oriented curriculum, has consequences on various aspects of learning, and on the determination of success criteria. If the previous curriculum is determined by the extent to which students' mastery of the lesson material, whereas in the competency-based curriculum, the emphasis on the mastered learning process has an impact on the changes in the behavior or performances of everyday students. According to Majid [1] curriculum 2004 or better known as competency-based curriculum:

... "Requires a learning process that lures creativity, has values, ethics, aesthetics, logic and kinesthetic, contextual, effectively and efficiently meaningful, and provide a meaningful learning experience.

Available Online: http://saspjournals.com/sjahss
LITERATURE REVIEW

The activity of analyzing the item is an activity that must be done to improve the quality of the problems that have been written. This activity is the process of collecting, summarizing, and using information from students’ answers to make decisions about each assessment [3]. The main objective of the item analysis in a test made by the teacher is to identify deficiencies in the test or in learning [4].

Based on this objective, the grain analysis activities have many benefits, among them are: First in order to help the test users in the evaluation of the test used. Second, it is highly relevant for the preparation of informal and local tests such as tests prepared by teachers for students in the classroom. Third, to support the writing of effective items. Fourth, the material can improve the test in the classroom, and the last is to improve the validity of the problem and reliability [4]. Linn and Gronlund [5] also add on the implementation of the item analysis activity which is usually designed to answer the following questions: 1) Is the question function correct? 2) Does this matter have the right level of difficulty? 3) Is the question free from irrelevant things? 4) Is the answer option effective?

The various descriptions above show that the item analysis is: to determine the problems that are defective or non-functional use, and to improve the quality of the item through the three components of analysis that is the level of difficulty, distinguishing power, and improve learning through the ambiguity of the problem.

Teachers other than as a teacher also act as evaluators. In the process of evaluation of education is needed the ability to analyze the problem, so the question used can reflect the ability of students. A good question will be able to evaluate the extent to which learners master the indicators that have been determined by the teacher. For that, the ability to analyze the problem after the test is needed by educators to evaluate whether the measuring tool used is not what is desired, among others can determine which students have or have not mastered the material taught by teachers and also can help improve test through revisions or remove ineffective questions, and to find out the diagnostic information to students whether they have / have not understood the material already taught [6].

Implementation of essential learning supervision is the provision of professional teacher assistance and improves the quality of learning in order to meet or exceed the standards set. That the responsibility of improving teacher professionalism and the success of learning is the obligation of the teacher itself. The effort given in the form of supervision of learning is more help and support from outside. The purpose of supervision according to Thomas [7]: 1. Growth of learners and improvement of education. 2. Supports educational leadership in ensuring the sustainability of education programs, 3. Cooperation to develop a fun atmosphere for learning. Thus supervision is carried out creatively constructively to encourage and create a conducive atmosphere for dynamic learning process.

Further described in the four categories of teachers expressed by Glickman [8] states based on the level of abstract thinking characterized by high cognitive abilities, imaginative, creative, adaptive, has a flexible teaching style, and the ability to overcome learning barriers and commitment levels characterized by a tendency to always actively involved with full responsibility. First is the professional teacher, the teacher who has a high level of abstract thinking and high commitment level. Second, the critical teacher, the teacher who has a high abstraction but low commitment. Third, teachers who are too busy, i.e. teachers who have a low abstraction but high commitment level. Fourth, less qualified teachers, where the teacher has a low level of abstraction and commitment level.

In conducting an evaluation of the program and implementation of supervision of learning should be presented first in this domain is the competence that should be owned by the supervisor to carry out the task of professional coaching teachers. There are at least three competencies according to Allan [9] who must be mastered by supervisors as a coach professionalism of teachers, namely (the nature of teaching) the nature of learning, (the nature of adult development) adult education approach, and (the characteristics of good and effective school) characteristics of advanced schools.

In general, the evaluation of the implementation of supervision of learning is related to the implementation of the supervision theory, which includes supervision planning, approaches and views used, the supervision of the teacher's teaching ability, and the supervision of the teacher's satisfaction and discipline. Supervision is a process of supervision over one's ability in achieving the success of the learning organization in accordance with the tasks it does [10]. The successful implementation of supervision is determined by the ability to analyze personal beliefs, attitudes, and values as the three components of philosophical principles of education.

Things that need to be done in the evaluation of learning supervision is the implementation of supervision of learning. In detail according to Jhon [10], things that need to be questioned include: 1) planning (teacher engagement, contract with teachers); 2) approach (scientific, artistic, clinical or intensive); 3)
supervision views used (directive, collaborative or non directive) based on teacher category. 4) Learning supplementary skills (conceptual, managerial, human, and technical) are applied proportionally, considering the implementation of such techniques, as well as the barriers to implementation; 5) the procedures undertaken (initial meeting, class observation and reverse meeting); 6) the satisfaction and discipline of teacher work.

RESEARCH METHODS

Location of Research
This research was conducted in “Masyarakat Damai” Junior High School Gunungsitoli, which was established since 1956, located in Madula Village Gunungsitoli District with distance from city 6 Km.

Research methods
The method used in this research is descriptive research method. In this study the researchers conducted data analysis by describing or describing the data that has been collected as it is without intending to make general conclusions or generalizations.

Research subject
Manuscript of the test made by the teacher analyzed in this research is the script of mathematics test since the academic year 2004/2005 - 2007/2008 while the process of assessing the results of cognitive aspect that is described is the result of the final study of the odd semester of the 2007/2008 academic year.

Research Instruments
To collect more accurate data in this study, researchers prepared several instruments in the form:

- Questionnaire of Test Preparation Steps. This questionnaire is in the form of a closed questionnaire, used to reveal how the steps taken by the teacher in the preparation of a set of tests, the process of scoring and follow-up done after the implementation of the exam block, midterm exam, and final exams semester by the teachers of mathematics in Swasta Masyarakat Damai Junior High School Gunungsitoli

- Document Analysis. The document that will be needed in this research is in the form of learning tools that have to do with the preparation of the test from the academic year 2004/2005 - 2007/2008 academic year.

- Interview. The implementation of the interview is to follow up the implementation of the questionnaire of the preparation of the test, the process of giving semester grade, the process of giving the final value of the odd semester of the academic year 2007/2008, and the follow-up done by the subject teachers after the execution of the test.

DATA ANALYSIS

Step-by-step Questionnaire
To analyze this questionnaire data using Guttman scale "if answer yes score 1 and if answered not score 0".

Validations and Reliability Test
To find out whether the mathematics test script that has been prepared by the teacher of mathematics in accordance with the material, construction and language domains, the researcher validates the triangulation to the teachers of mathematics subjects as much as 2 (two) persons, and 1 (one) language teacher Indonesia, where each test item consists of 2 (two) columns. Column 1: if yes score 1, otherwise score 0 and processed using Guttman scale. Column 2 with rating scale: 1 = invalid; 2 = less valid; 3 = quite valid; and 4 = valid.

Test the validity of the test used Product Moment correlation formula as follows:

\[
r = \frac{N \sum XY - (\sum X)(\sum Y)}{\sqrt{(N \sum X^2 - (\sum X)^2)(N \sum Y^2 - (\sum Y)^2)}}
\]

Where:
- \( r_{xy} \) = The coefficient of validity between variables \( x \) and \( y \)
- \( N \) = Number of test participants
- \( \Sigma X \) = Number of scores per test item
- \( \Sigma Y \) = Total total score

Test Reliability test, for multiple choice questions used the formula KR-20:

\[
r = \frac{k}{k - 1}\left(1 - \frac{\sum pq}{\sigma^2}\right)
\]

Where:
- \( p \) = correct proportion of answers
- \( q \) = the proportion of incorrect answers
- \( k \) = many test items
- \( \sigma^2 \) = score variance

For Testing Level of Tests Tests used the formula:

\[
TK = \frac{B}{N}
\]

Where:
- \( TK \) = Problem difficulty level item,
Different Power Test

The test is used:

$$DP = \frac{2(BA - BB)}{N}$$

Where:
- $DP$ = Power difference test
- $BA$ = Number of correct answers in the top group
- $BB$ = Number of correct answers in the lower group
- $N$ = Number of students who do the test

Distactor Function Analysis

The objective test of multiple-choice forms in which each item issued in a learning result test has been completed with several possible answers or often known as optional or alternative terms. Option or alternate amounts ranging from three to five, and from the possible answers attached to each item item, one of them is the correct answer (key answer), while the rest is the wrong answer. Wrong answers are commonly known as distractors. The distractor is said to have performed its function properly if the distractor has at least been selected by 5% of all test takers.

### FINDING AND DISCUSSIONS

In this section we will describe the questionnaire and interview data from 2 (two) mathematics subject teachers (IATM and RH) with the following results:

#### Test Preparation Steps

<table>
<thead>
<tr>
<th>Respondent</th>
<th>Test preparation steps</th>
<th>Reproducible</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>IATM</td>
<td>Yes 6, No 2</td>
<td>0.67</td>
<td>Not Accepted</td>
</tr>
<tr>
<td>RH</td>
<td>Yes 6, No 2</td>
<td>0.67</td>
<td>Not Accepted</td>
</tr>
</tbody>
</table>

From Table 1, out of the 8 expected steps in the compilation of the question turned out to be only 6 steps done by IATM and RH. Calculations by Guttman scale are 0.67 with an unacceptable category reproduction. The steps to prepare the questions that have been done by IATM are: to choose the type of test according to the competence to be measured, make the problem yourself, correct or evaluate the result of multiple choice test according to the predetermined instruction, correct or evaluate the result of the essay test every item in accordance with the score, the weight that has been set before, perform analysis of student learning outcome test at each written examination, can analyze the written test of student learning outcomes. The steps of preparing the questions that have been done by RH are: choosing the type of test in accordance with the competence to be measured, making your own question, choosing the type of test in accordance with the predetermined instruction, correct or evaluate the result of multiple choice test results in accordance with the predefined instructions, students at each written exam, can analyze the written test of student learning outcomes. The steps for preparing the questions that have not been implemented by RH are: to prepare the test grille for each evaluation of learning outcomes and to correct or assess the results of the essay test every item in accordance with the score, the predetermined weight.

#### Test Validity

The validity of teacher-made tests under study consisted of 24 test packets with the following details:

a. 2004/2005 Lessons Year, which consists of one class of Class VII with the title of the test of Block 1, Block 2, and Block 3 and its implementation in accordance with the principal program three times in one semester.

b. Year of Study 2005/2006, which consists of:
   1. Class VII, with the title of Block 1, Block 2 and Block 3.
   2. Class VIII, with the title of Block 1, Block 2 and Block 3 and its implementation in accordance with the principal’s work program three times in one semester.
   c. Lesson 2006/2007, which consists of:
      1. Class VII, with the title of Block 1, Block 2 and Block 3.
      2. Class VIII, with the title of Block 1, Block 2 and Block 3.
      3. Class IX, with the title of Block 1, Block 2 and Block 3, and its implementation in accordance with the principal’s work program three times in one semester.
   d. Lesson 2007/2008, which consists of:
      1. Class VII, with the title of midterm exam, and final exam of semester with test material consist of some subject matter which has been programmed by subject teacher and its implementation according to principal work program.
      2. Class VIII, with the title of midterm exam, and final semester test script with test material consist of some subject matter which has been programmed by subject teacher and its implementation in accordance with principal work program.
      3. Class IX, with the title of midterm exam, and final semester test script with test material consist of some subject matter which has been programmed by subject teacher and its implementation in accordance with principal work program.
To investigate the validity of tests on the material domain, the construction domain, and the language domain all test packages have been validated to experienced math teachers (achievement teachers), with results as shown in the following table:

### Table 2: Recapitulation of Test Validation

<table>
<thead>
<tr>
<th>Academic Years</th>
<th>Class</th>
<th>UB/UTSG/UASG</th>
<th>Reproducible Accepted (%)</th>
<th>Reproducible Not Accepted (%)</th>
<th>Not Valid (%)</th>
<th>Less Valid (%)</th>
<th>Quite Valid (%)</th>
<th>Valid (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004/2005</td>
<td>VII</td>
<td>UB 1</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UB 2</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UB 3</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2005/2006</td>
<td>VII</td>
<td>UB 1</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UB 2</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UB 3</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2006/2007</td>
<td>VIII</td>
<td>UB 1</td>
<td>0</td>
<td>100</td>
<td>20</td>
<td>80</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UB 2</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UB 3</td>
<td>0</td>
<td>100</td>
<td>80</td>
<td>20</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2007/2008</td>
<td>IX</td>
<td>UB 1</td>
<td>0</td>
<td>100</td>
<td>40</td>
<td>60</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UB 2</td>
<td>0</td>
<td>100</td>
<td>95</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UB 3</td>
<td>0</td>
<td>100</td>
<td>90</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UTSG</td>
<td>35</td>
<td>65</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>VII</td>
<td>UASG</td>
<td>55,55</td>
<td>44,44</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>VIII</td>
<td>UASG</td>
<td>45</td>
<td>55</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>IX</td>
<td>UASG</td>
<td>10</td>
<td>90</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>Average (%)</td>
<td></td>
<td></td>
<td>10.93</td>
<td>89.07</td>
<td>16.04</td>
<td>55.63</td>
<td>28.33</td>
<td>0</td>
</tr>
</tbody>
</table>

Notes: UB = Block Test; UTSG = Mid-Test; UASG = Final Test

a. The average reproductive level of acceptable test items from the 2004/2005 school year to the 2007/2008 school year was 10.93% (49 items) while the test items were not accepted 89.07% (395 items), by scale Guttman.

b. The average rating scale is: 16.04% (71 item) valid test items, 55.63% (247 item) valid test items, and 28.53% (126 item) valid test items, as well as valid 0% test items.

To determine the validity of test items, the researcher has feasibility test with the result of recapitulation of test item validity as in Table 3 below:

### Table 3: Table Validity and Reliability Test Recapitulation

<table>
<thead>
<tr>
<th>Academic Years</th>
<th>Class</th>
<th>UB/UTSG/UASG</th>
<th>Valid %</th>
<th>Not Valid %</th>
<th>Reliable %</th>
<th>Not Reliable %</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004/2005</td>
<td>VII</td>
<td>UB 1</td>
<td>75</td>
<td>25</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UB 2</td>
<td>45</td>
<td>55</td>
<td>R</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UB 3</td>
<td>35</td>
<td>65</td>
<td>R</td>
<td>-</td>
</tr>
<tr>
<td>2005/2006</td>
<td>VII</td>
<td>UB 1</td>
<td>25</td>
<td>75</td>
<td>R</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UB 2</td>
<td>15</td>
<td>85</td>
<td>R</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UB 3</td>
<td>30</td>
<td>70</td>
<td>-</td>
<td>TR</td>
</tr>
<tr>
<td>2006/2007</td>
<td>VIII</td>
<td>UB 1</td>
<td>30</td>
<td>70</td>
<td>R</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UB 2</td>
<td>40</td>
<td>60</td>
<td>-</td>
<td>TR</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UB 3</td>
<td>35</td>
<td>65</td>
<td>-</td>
<td>TR</td>
</tr>
</tbody>
</table>
From Table 3 above it turns out that the average of valid multiple-choice test items is 42.23%, the invalid test item is 57.77% (256 items); while the valid test item test essay is 70%, the invalid test item is 30%.

- Reliability Test. From Table 9, obtained a reliable 12 test packs (54.55%), which is unreliable are 10 test packs (45.55%); a reliable test of 2 test packages (40%), which is unreliable is 3 test packs (60%).

**Table-3: Recapitulation of Exchange Index, Differentiating Power and Test Recipient**

<table>
<thead>
<tr>
<th>Academic Year</th>
<th>Class</th>
<th>UB/USTG/UAGS</th>
<th>Exchange Index</th>
<th>Power Differentiation</th>
<th>Spieler</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Easy (%), Medium (%), Difficult (%)</td>
<td>(%), (%), (2)</td>
<td>(%), (%), (%), (2)</td>
<td>(%), (%), (2)</td>
</tr>
<tr>
<td>2004/2005</td>
<td>VII</td>
<td>UB 1 10, 65, 25</td>
<td>25, 20, 25</td>
<td>30, 90, 10</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UB 2 45, 30, 15</td>
<td>20, 20, 25</td>
<td>25, 45, 45</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UB 3 50, 45, 5</td>
<td>15, 10, 5</td>
<td>70, 25, 75</td>
<td>0</td>
</tr>
<tr>
<td>2005/2006</td>
<td>VII</td>
<td>UB 1 25, 55, 20</td>
<td>15, 5, 20</td>
<td>60, 50, 50</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UB 2 50, 5, 35</td>
<td>10, 5, 10</td>
<td>65, 25, 65</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UB 3 15, 35, 50</td>
<td>15, 0, 10</td>
<td>75, 65, 35</td>
<td>0</td>
</tr>
<tr>
<td>2006/2007</td>
<td>VIII</td>
<td>UB 1 15, 30, 55</td>
<td>0, 5, 40</td>
<td>55, 55, 45</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UB 2 5, 45, 50</td>
<td>7, 0, 0</td>
<td>65, 10, 90</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UB 3 5, 20, 75</td>
<td>10, 15, 10</td>
<td>65, 90, 10</td>
<td>0</td>
</tr>
<tr>
<td>2007/2008</td>
<td>VII</td>
<td>UB 1 20, 80, 0</td>
<td>40, 20, 20</td>
<td>20, -</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UB 2 0, 40, 60</td>
<td>0, 5, 15</td>
<td>80, 90, 10</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UB 3 0, 45, 55</td>
<td>10, 10, 15</td>
<td>65, 70, 30</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>VIII</td>
<td>UB 1 20, 80, 0</td>
<td>20, 0, 0</td>
<td>80, -</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UB 2 0, 35, 65</td>
<td>35, 20, 5</td>
<td>40, 80, 20</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UB 3 0, 50, 50</td>
<td>10, 10, 20</td>
<td>60, 45, 55</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>IX</td>
<td>UB 1 50, 15, 35</td>
<td>10, 5, 25</td>
<td>60, 55, 45</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UB 2 60, 20, 20</td>
<td>5, 15, 20</td>
<td>60, 30, 70</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UB 3 60, 30, 10</td>
<td>10, 10, 25</td>
<td>55, 5, 95</td>
<td>0</td>
</tr>
<tr>
<td>2007/2008</td>
<td>VII</td>
<td>UTSG 0, 60, 35</td>
<td>0, 5, 30</td>
<td>60, 95, 0</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UASG 5.56, 27.78, 66.67</td>
<td>16.67, 5.56, 16.67</td>
<td>61.11, 46.67, 53.33</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>VIII</td>
<td>UTSG 10, 55, 35</td>
<td>10, 20, 25</td>
<td>45, 85, 15</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UASG 0, 61.11, 38.89</td>
<td>22.22, 5.56, 16.67</td>
<td>55.56, 73.33, 26.67</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>IX</td>
<td>UTSG 30, 45, 10</td>
<td>0, 20, 25</td>
<td>40, 70, 15</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UASG 22.22, 55.56, 22.22</td>
<td>11.11, 22.22, 16.67</td>
<td>50, 53.33, 46.67</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Average % 24.49, 42.89, 32.62</td>
<td>13.21, 10.56, 17.29</td>
<td>58.94, 56.97, 43.03</td>
<td>2.5</td>
</tr>
</tbody>
</table>
From Table 4 above it turns out, out of 24 test packages with total item number 444 consisting of: multiple choice 425 item, and item 19 essay, it turns out:

a. Test difficulty index:
   - Easy = 24.49% (109 item)
   - Medium = 42.89% (190 items)
   - Difficult = 32.62% (145 item)

b. Power difference test:
   - Problem received either = 13.21% (59 items)
   - Problem received but needs to be improved = 10.56% (47 items)
   - Problem fixed = 17.29% (77 item)
   - Problem not used / discarded = 58.94% (261 items)

c. Spieler:
   - Effective = 56.97% (242 items)
   - Not Effective = 43.03% (183 items).

**DISCUSSION**

Based on the results obtained through questionnaires and interviews from both IATM and RH teachers on:

**Step-by-Step Preparation Steps**

The result of questionnaire from Table 2, it turns out:

a. IATM there are 2 (two) steps that have not been implemented on the compilation of questions, namely:
   1). Develop a test grille for each evaluation of learning outcomes.
      The problems prepared by each teacher to produce material / test valid and reliable, and then should be designed in the form of matrix test lattice. This happens that the IATM has not understood the meaning of the test grille, the way the test lattice is made (never following the instructional exercises) or the lack of control of the principal's leadership elements so that the IATM does not make the test grille of each evaluation of the learning outcomes. As a result, the findings of the calculation of validity and reliability of the test are low.
   2). Taking questions from books (problem banks).
      The problem of the book (question bank) is a standardized test script that has been tested before it is summarized in a book that has met the good test criteria that is valid and reliable.
      This happens that IATM does not have a book (problem bank), because IATM does not take part of the book (question bank) resulting in test items tested to less qualified learners. This also affects students who are achievers according to the school level, if they follow the competition outside school or continue at higher levels because of lack of experience on standardized test items resulting in many unsuccessful.

b. RH there are also 2 (two) steps that have not been implemented on the preparation of the problem, namely:
   1). Develop a test grille for each evaluation of learning outcomes.
   2). Correct or assess the results of the essay test each item in accordance with the score, the weight that has been set before.

The basis for determining the score is based on many problems as well as the level of difficulty of the problem, while the weight of the question is based on the depth and extent of the material in question or the level of complexity or complexity of answers demanded by a problem.

This happens, RH lacks understanding of the basic determination of the score and weight of the essay test and the evaluation procedure of the essay test and the RH is afraid to be said to fail by the parents of the students / co-teachers. As a result of RH is not based on reference (score, and weight) that has been determined previously resulted in the value of student learning outcomes are not in accordance with student achievement itself.

The interview result of the preparatory steps, IATM and RH deviate from the truth. It should be that the problem prepared by a teacher to produce valid and reliable test materials should be designed in the form of matrix test lattice. These steps are: determining the subject matter tested, preparing the test grille, writing the problem, examining the problem, assembling the matter into a test device, and scoring it. IATM and RH ignore these steps because they feel senior or have experience in composing the problem. As a result of not paying attention to these steps resulted as the result of finding a lot of test items that are not valid and not reliable (test items less quality).

**Validity of Master's Artificial Test Manuscripts**

Based on the results from Table 8, on the recapitulation of the validity of the odd semester end test items from the 2004/2005 academic year to the 2007/2008 academic year, according to Guttman's scale the reproducibility level of the test items in terms of the material, construction, and language aspects is accepted 10, 93% (49 items) and 89.07% (395 items) were not accepted. The result of the average scale of the invalid rating is 16.04% (71 item) meaning the question can not be used, less valid 55.63% (247 items), meaning the question can be used but needs improvement, and quite valid 28.53% (126 items), meaning the question can be used but need consultation and valid 0%.

Based on the results of calculating the validity of the test items in Table 9 consisting of 24 test packages (444 items) of valid multiple choice 42.23%
(188 items), meaning good item item, and 57.77% (256 item) means the problem can not be used. A valid 70% test item (13 items) means good item point, and 30% (6 item) is invalid, meaning it cannot be used. Specifically, an invalid grain test means that it can not be used, meaning that the test item does not match what the tested intends to measure.

Reliability Test

Based on the results of the calculation of the validity of the test in Table 9, it turns out that from the 22 test packages of 12 multiple test options (54.55%), reliable means that can be accepted as a means of measuring the students' learning ability that can provide accurate; while 10 test packs (45.5%) can not be used or discarded, while the test forms the essays of 5 test packets, reliable 2 test packs (40%), and unreliable 3 test packs (60%). Unreliable test packets are refused or partially removed and can not be used as question banks.

Validity and Reliability tests

From the result of the finding of validity of multiple choice test items that belong to valid category is 42.23% and reliable test reliability is 54.55% and the test essay, which is valid category, is 70% and reliable test is 40%. The above mentioned occurred because IATM and RH in preparing / preparation of exam questions not based on the test grid (according to data on the implementation of the questionnaire). As a result of IATM and RH in preparing the preparation of the test item not based on the test grid resulted in the script of the odd semester test of the academic year 2004/2005 up to the 2007/2008 academic year which belongs to the multiple choice valid category is 42.23%, and reliable is 54.55% while the test of valid category is 70% and reliable is 40%.

From the findings of valutators and further researchers is concerned about: the incompatibility between the realm of material, the construction and language aspects, and not the right answer key, no key answers, not consecutive options, not good or not working, and sentences about not precise or dubious students who want to be asked, as well as the formulation of sentences about which is not clear.

CONCLUSIONS, IMPLICATIONS, AND SUGGESTIONS

CONCLUSION

The steps of the compilation of the problems pursued by the teacher of mathematics subjects not yet in accordance with the steps of preparing the real question. The process of scoring the results of learning by teachers on the final exams of the semester has not guided the reference that has been agreed. Teachers mathematics subjects have not done follow-up in the form of remedial and enrichment well. The validity of teacher-made tests in terms of the material domain, the construction sphere, and the language domain is only 10.93%, which belongs to the accepted category according to Guttman's scale. From the findings of the validity of the multiple choice test items that belong to valid category is 42.23% and the test essay, which is valid category, is 70%. The reliability of multiple-choice tests including the reliable category is 54.55% and the reliable test essay is 40%.

Implications

Step-by-step preparation that is not based on the test grid can lead to the test item is not valid and not reliable. Procedure of scoring that is not based on predefined reference (score and weight) can result in the value of student learning outcomes are not in accordance with the achievements achieved.

Suggestion

The math teacher in preparing the script should be based on the test grid. Each student's value of the students' learning outcomes is based on a predefined reference and follows the procedure of examining the answer sheets, in order to ensure the credibility of the quality of the graduates. The implementation of remedial programs for students who do not achieve the Minimum Criteria and enrichment programs for students who score ≥ 85 are performed outside of the effective learning hours. At the end of each examination, the items are analyzed the problem level, the differentiation of the problem, the effectiveness of the deceiver and the validity and reliability of the matter in order to know the quality of the problem.

REFERENCES