THE DEVELOPMENT OF UAS FOR JUNIOR HIGH
BY MGMP BAHASA INGGRIS KABUPATEN PASURUAN

Maharani Kusumawardani Mas’ud
Email: queenrich@rocketmail.com
State University of Malang

ABSTRACT: This study is intended to evaluate the UAS made by MGMP Bahasa
Inggris Kabupaten Pasuruan in terms of the stages of test development, the design
of the scoring method, and the multiple-choice question item construction. The
design of this study was a qualitative case study. The study yields two findings.
First, the stages of test development carried out by the test development team are
good and mostly follow the expert guidelines. Second, the design of the scoring
method is appropriate for the type of the test. Third, the multiple-choice questions
are not well-constructed.

Keywords: test development, scoring method, multiple-choice questions, MGMP

Testing is a technique to obtain information (Hopkins and Stanley, 1981). Most
language tests measure one’s ability to perform language, that is, to speak, write, read,
or listen to a subset of language. According to Mayo (1980), some important reasons
tests are administered in today’s schools are to provide diagnostic information about
individuals and group performance, assess the learning progress of individuals and
groups, predict future academic performance, and evaluate the effectiveness of
instruction.

At the end of a semester, teachers often give students an achievement test.
According to Tinambunan (1988:27), an achievement test should measure clearly
defined learning outcomes that are in harmony with the instructional objectives. The
test that can measure the intended learning outcome must have a good quality.
Bachman (2003:18) states that in developing language tests, we must take into
account considerations and follow procedures that are characteristic of tests and
measurement in the social sciences. Teachers can have a good test by developing it
through proper stages.

According to Sulistyow (2002:42-48), test development is a thorough and
systematic procedure to follow in making a set of test. There are consecutive standard
procedures that a test developer needs to follow in order to develop a good test. It
starts with the formulation of the test objectives. Then, competences to be measured
are derived from the competence standards and basic competences. The next step is
the construction of blueprints. Then, the blueprints need to be reviewed. The next step
is item writing. The items already written are then assembled into a set of test. When
the test items are arranged in the form of a test consisting of sub tests or test items,
then it comes to the process of expert review. Following this stage, there are two other possible steps that may be taken: putting the assembled items in the item bank in which the tests are kept before it is used, or using the test as planned. After that, the test can be administered. Then there comes another important process called test analysis. The outcome of the test analysis reflects the quality of the test.

The results of language tests are most often reported as numbers or scores, and it is these scores, ultimately, that test users will make use of (Bachman and Palmer, 1996, 193). The scoring methods used to arrive at scores must be decided earlier before the test is held. Alderson, Clapham, and Wall (1995: 148) state:

Decisions need to be taken on whether simply to add marks up to arrive at a total score for the test, or whether to give some items more importance than others. Testers need to decide which candidates can be considered to have performed adequately, and thus to have passed the test and which have failed.

Each item in a test must be well-constructed based on the type of the items. For multiple-choice question type, the criteria are every item must use clear and simple language, measure only one formulated problem, use proper grammar and, avoid any form that might provide clues for the test takers (Gronlund and Waugh, 2009:93-106).

A unique phenomenon is analyzed in this study. It is the semester test, which is called as Ujian Akhir Semester (henceforth, UAS), in Kabupaten Pasuruan. Based on the Regulation of Ministry of National Education Number 20 year 2007 on July 11th, 2007 on Education Assessment Standard, it is the activity carried out by teachers to measure students’ achievement to the competence at the end of a semester. The UAS is made by a test development team under a teacher association, the Musyawarah Guru Mata Pelajaran (henceforth MGMP), Kabupaten Pasuruan. This test is used in all schools in the area. Usually, every teacher develops their own test for their own students for UAS because the teachers are the ones who know what and to what extent of materials have been given to students. Based on the Board of Educational National Standard, as stated in Panduan Penyusunan Kurikulum Tingkat Satuan Pendidikan Dasar dan Menengah (2006:5), Kurikulum Tingkat Satuan Pendidikan (henceforth KTSP), the recent curriculum when this study was conducted, is an implementation of educational reformation in Indonesia that gives autonomy to schools to develop their own curriculum based on their potential and needs.

Previous studies related to test development have not raised this phenomenon. The first is a survey conducted by Lestari (2010) about the implementation of authentic
assessment conducted in classrooms by public junior high schools English teachers in Malang Municipality. The findings showed that the teachers were not well-informed about the implementation and techniques of authentic assessment and other kinds of assessment. The second is Widowati’s analysis (2011) of the teacher-made try-out test. The research aimed to analyze the quality of the test items in terms of the validity and reliability. Based on the research findings, it was found that the teacher-made try-out test of UAN 2010/2011 of junior high schools in Malang has good face validity, content validity, reliability, item validity, discrimination index, and item difficulty. However, the research found out that the distracters of the items are not good. Next, it is Ratnafuri’s (2011) study with the item analysis on the teacher-made test for the final exam of the semester. Similar to what Widowati did, she analyzed the test in terms of the validity and reliability. The findings of this study show that the test has low quality because it has moderate test construction, moderate content validity, moderate reliability, low index of difficulty, low index of discrimination, and many ineffective distracters although it has high practicality. The last is Prastiwi’s study (2011) on the test development of a teacher-made test in SMAN 3 Malang. Her research was about the teachers’ knowledge of test development, the stages that the teachers followed in developing a test and the obstacles and supporting factors in developing a test. The findings show that the English teachers in SMAN 3 Malang lacked the knowledge of stages in test development, the stages of test development by the English teachers did not follow the standard procedure in test development, and the obstacles or problems faced by the English teachers in the test development were mostly due to the limitation of time. Unlike the previous studies, in this study, the test cannot be considered as a teacher-made test since it is used by many schools. Therefore, it must be well-constructed.

This study aims at investigating how well MGMP Bahasa Inggris Kabupaten Pasuruan develop UAS of the first semester of 2012/2013 of junior high schools in Kabupaten Pasuruan. The development includes the stages of test development, scoring methods, and construction of the multiple-choice questions.

METHOD

This study is a qualitative study because the purpose of this study is to understand and describe the phenomenon happening in the subjects being observed in natural contexts. The subjects of this study are four teachers who are the members of the test development team for UAS Bahasa Inggris. The first subject is the leader of MGMP Bahasa Inggris Kabupaten Pasuruan. The second subject is the test maker of
the seventh and eighth graders. The third subject is one of the test makers of the ninth graders. The fourth subject is one of the teachers who are in charge of the test verification. All of the data were obtained through interviews and documents collected from the teachers. The interview was conducted with an interview guide. The document used in this study is the UAS Bahasa Inggris of the seventh grade of the first semester of 2012/2013. The data from the interviews were compared to the guideline of the test development process adapted from Sulistyanto (2002) and ALTE (2011) and the guideline of scoring methods adapted from Alderson, Clapham, and Wall (1995), Bachman and Palmer (1996), and (Popham, 1981). The data from the document were used to analyze the construction of the multiple-choice questions. The guideline was adapted from Gronlund and Waugh (2009:93-106), Wood, Merrill, Sudweeks, and Burton (1991:10-13), and Lindvall and Nitko (1975: 58-59).

A qualitative analysis was employed in order to answer the research problems of this study (McMillan, 1992:479). This involves a systematic process of selecting, categorizing, comparing, synthesizing and interpreting data to provide explanations of the single phenomenon of interest. First, in the selecting process, the data collected from interviews, which were recorded, were attended and transferred into transcriptions. Then, the data were reduced and selected. The relevant data were used while the irrelevant data were omitted. For the document, the researcher selected the test for the seventh grade. The test was in a hard copy so it must be scanned to ease the analysis. The relevant data from interviews, then, were arranged into systematic summaries of the information needed. The data from the document were evaluated based on the criteria of good multiple-choice construction. For the data from both the interviews and the document, the researcher made tabulations. This was the process of categorizing. After that, the data from the interviews were compiled to complete the information needed. Then, the information was compared to the guidelines of test development and scoring methods. Meanwhile, data from the document were compared to the guidelines of multiple-choice item construction. This was the comparing process. Next, all of the data were brought together based on the result of the comparing process. This was the process of synthesizing. Then, the result from the data analysis was interpreted to answer each research problems.

FINDINGS AND DISCUSSIONS

Based on the result of the interviews, the test development team is divided into three groups. The first group takes care of the test purpose recognition. The second
group is the test makers and the third group is the test verification team. The findings of this study are divided into three major findings: the stages of test development, scoring methods, and construction of multiple-choice questions.

**Stages of Test Development**

The test development team of *MGMP Bahasa Inggris Kabupaten Pasuruan* did not follow the recommended stages of test development for the *UAS*. They had a unique order of stages that they thought were the most appropriate for them. However, they missed one stage that is mentioned in the guideline—the try-out stage.

The first stage was the test purpose-recognition stage. The activities carried out in this stage were incomplete. Downing and Haladyna (2006) suggest the activities that must be carried out in the first stage. They are constructing desired test interpretation, the format of the test, and the major sources of validity evidence; determining clear purposes, desired inferences, psychometric model, and timelines, and deciding ways to keep security and control quality. Unfortunately, the test development team did not consider the major sources of validity evidence. Neither did they prepare the test interpretation and the psychometric model. It is because in this stage they decided to limit their responsibility up to the test assembling, excluding the scoring and the test analysis mentioned in the guidelines. That is why the researcher said that this stage is incomplete.

The second stage was the making of blueprints. The stakeholder informed the team that they had to develop the test. The test makers made the blueprint, item-cards, and the live test materials at home. In this stage, the test developer had to determine, at least, the topic to be covered, the domain to be examined, the level of difficulty to be considered, and, more importantly, the number of the test items (Sulistyo, 2002). The test-makers team did all these activities. Furthermore, the blueprint was one of the documents they had to submit to the Department of Education. In the blueprint, they determined the Competence Standard and Basic Competence (*SK-KD*) for the seventh and eighth grades and the Graduate Competence Standard (*SKL*) for the ninth grade, the number of items, and the level of difficulty. The level of difficulty was determined by some test analyses carried out by the test makers for the seventh and eighth grades. Although much of the information was determined in the meeting with the leader of the test development team, the test makers were involved in the decision-making. This is in line with ALTE (2011) who state that much of this
information must come from the sponsor of the test. The test makers’ job, then, is to determine the objectives, materials, and level of difficulty.

After making the blueprints, the test makers developed the item-cards. This is in line with Sulistyö’s statement (2002) that this stage involves the writing of items on test-item cards. The purpose is to show the specifications of each item. They did not review the item-cards because that is the job of the verification team. After that, the test makers assembled the items. Sulistyö (2002) suggested that in this process test makers take several considerations. They should order the test items from easy to difficult ones, plan the estimation of test reliability, group test items under certain sub tests, write test instructions, and set the test layout. Not all considerations taken by the test makers are in line with the guidelines. The emphasis was on the layout setting. The rest of the assembling activities were done by somebody in the Department of Education in Kabupaten Pasuruan. Even though the test makers did not finish the test assembling, they already had an agreement with this person about the layout of the test. Besides, they sent him not only the print out of the test but also the softcopy to make sure that the test would be printed as the test makers had developed and test verification team had approved.

On the due date of collecting the test makers’ work, they held a meeting where test makers and test reviewers reviewed the test makers’ work. The procedure of the review is as follows. First, they checked the test makers’ work on a one reviewer–one grade basis. Then, they checked different grades. Any points that had to be corrected, would be discussed with the test makers. The verification team revised them, whenever necessary, based on the result of discussion with the test makers.

Both Sulistyö (2002) and ALTE (2011) suggest asking experts to review each of them. Although there is a special team for verification, the members are not actual experts in test construction. Therefore, the reviews are peer reviews. However, it does not mean that this is not in line with the guidelines. Sulistyö (2002) state that in the case where an expert is not available, peer review would be sufficient for the purpose. According to ALTE (2011), information for reviews can be gathered in a number of ways, such as piloting (asking some test takers to sit the test) and the analysis of the responses; consulting colleagues (peer reviews); and consulting other stakeholders. The existence of this test verification team is in line with ALTE (2011) because it supports the most important rule in item reviews: each item should not be checked by the item writers themselves.

After that, the test development team gave the final products (blueprints, item-
cards, answer key, and live test materials) to the Department of Education in Kabupaten Pasuruan. The process of assembling the test, then, was continued by a team in the Educational Department. They would add a letterhead, manage the layout, decide the material for the paper, and send the test to a printing company or publisher.

In short, there is only one stage that was missed which was the try out test stage. The next point that is different from the guidelines is the review stages. In the guideline, the review should be conducted after each stage is finished. It means that the blueprint review should be carried out after the making of blueprint; items review should be performed after item writing; and assembled test review should be processed after the test assembling. The regulation taken by the test development team, however, is different. All the reviews were done in one stage.

**Scoring Method**

Based on the findings, the scoring methods were not well designed. The focus of the test development team was mostly on the making of the items. That is why they designed the test in such a way that the scoring is the easiest and most practical. It is proved by the absence of the rater team. The approach they used was the number of items successfully completed. It is appropriate for scoring multiple-choice questions because they are categorized as a selected response type and it is in line with the guidelines. Furthermore, the use of rating scale is impossible because it is typically used with tests that require test takers to produce an extended response (Alderson, Clapham, and Wall, 1995). It is supported by Bachman and Palmer (1996) who state that scoring as the number of items successfully completed can be used to measure specific areas of language knowledge with the assumption that the test was sufficiently well defined by the way the item was designed and written.

Hence, the procedure in scoring the test was number-right scoring or dichotomous. In the absence of a rater team, it is the most practical way because the test development team do not need to include a different weight or credit to each item in the scoring guide. However, according to Frary (1989), the weakness of this procedure is that this is advantageous for test takers to answer every item in the test. It is because even if they choose some alternatives at random, they will not get any penalty for wrong answers and, if their answers are right, it does not fairly measure the intended learning outcome.

Based on the findings, the test did not use any weighting because they did not mean that some items are more important than the others are. However, the multiple-
choice questions used in the test are in several varieties such as arranging sentences, reading comprehension, grammar and vocabulary knowledge, and cloze test. It makes possible that some items will require more advanced proficiency and take more time to answer (Alderson, Clapham, and Wall, 1995). For example, the reading comprehension task is more difficult than the sentence-arranging task. The two possibilities are the reason why weighting is necessary. Moreover, Alderson, Clapham, and Wall (1995) also mention that items believed to be more central to the curriculum or to the concept of proficiency are necessary to get more weight than the others are. However, the teachers’ knowledge about weighting is limited so the absence of weighting is understandable although this is not the most appropriate method based on the guidelines.

Next, the English teachers scored the test. It means there was no rater team. Hence, after the test was administered, the students’ works were scored by their own English teachers right away based on the answer key. Some schools, however, have equipped themselves with a scanner. These schools usually decided to utilize their scanner. Therefore, they must produce computer answer sheets themselves and the answer sheets given by the Department of Education in Kabupaten Pasuruan were discarded. This is in line with Gronlund (1985) who state that in scoring tests the aim is to obtain accurate results as rapidly and economically as possible. Nonetheless, without control of the test provider, it cannot be guaranteed that the students’ results will be scored objectively although the answer key is already available. The test development team argued that the current curriculum (KTSP) requires schools to be independent in managing the curriculum. This is a bit confusing because administering a test by using this regulation is in contrast with independency.

According to the teachers, they prepared scoring guides to be distributed to the schools. The scoring guide only included the key answer. The test development team did not provide any briefing for the English teachers because the scoring method is very easy to do and do not require any special skill. Unfortunately, the Department of Education in Kabupaten Pasuruan did not distribute the answer key to the teachers in fear that it would spread to students before the test was carried out. Consequently, the teachers had to make the answer key themselves.

**Construction of the Multiple-choice Questions**

The results of the analysis on the construction of the UAS of the first semester of 2012/2013 of junior high schools in Kabupaten Pasuruan made by MGMP Bahasa
Inggris are presented in Tables 1 and 2.
Table 1 The Results of the Analysis of the Multiple-Choice Question Item Construction

<table>
<thead>
<tr>
<th>Fulfillment of Principle</th>
<th>Frequency</th>
<th>%</th>
<th>Item number</th>
</tr>
</thead>
</table>
Table 1 shows that the teachers generally follow the principle of constructing multiple-choice question items since no item is categorized as fair. Even, an item with the most mistakes, which is 2% from the total items, are categorized as good. There are 26% percent of the total items that fulfilled the principles completely.

Table 2  The Distribution of Mistakes Made by the Teachers in Constructing the Multiple Choice Question Items

<table>
<thead>
<tr>
<th>No.</th>
<th>Criteria</th>
<th>Frequency</th>
<th>%</th>
<th>Items no.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Using of grammar, punctuation, and spelling improperly</td>
<td>27</td>
<td>50%</td>
<td>1, 2, 3, 4, 5, 6, 7, 8, 13, 16, 17, 18, 20, 23, 24, 27, 28, 29, 30, 32, 37, 38, 39, 44, 46, 47, and 48.</td>
</tr>
<tr>
<td>2</td>
<td>Unclear directions</td>
<td>6</td>
<td>11.11%</td>
<td>31, 37, 41, 42, 47, and 48.</td>
</tr>
<tr>
<td>3</td>
<td>changing pages in the middle of item</td>
<td>2</td>
<td>3.7%</td>
<td>35 and 36.</td>
</tr>
<tr>
<td>4</td>
<td>Assessing more than one formulated problems</td>
<td>2</td>
<td>3.7%</td>
<td>4 and 46.</td>
</tr>
<tr>
<td>5</td>
<td>Unclear, and confusing stem</td>
<td>1</td>
<td>1.85%</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>Using negative wording without emphasizing</td>
<td>0</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>7</td>
<td>Repeating some words in options</td>
<td>8</td>
<td>14.81%</td>
<td>5, 6, 10, 16, 19, 39, 43 and 45.</td>
</tr>
<tr>
<td>8</td>
<td>Not homogeneous content of alternatives</td>
<td>4</td>
<td>7.41%</td>
<td>11, 26, 27, and 46.</td>
</tr>
<tr>
<td>9</td>
<td>Undue stated alternatives</td>
<td>0</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>10</td>
<td>Grammatically inconsistent</td>
<td>0</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>11</td>
<td>Using length as the clue</td>
<td>4</td>
<td>7.41%</td>
<td>6, 11, 13, and 16</td>
</tr>
<tr>
<td>12</td>
<td>More than one accepted answers</td>
<td>0</td>
<td>0</td>
<td>-</td>
</tr>
</tbody>
</table>
Table 2 shows that the test maker makes fifty-four mistakes in total. The most frequent mistake concerns with the first principle, that is, the test maker was not aware of the punctuation and grammar. The total of the mistakes they made is fifty-four.

A good item must use consistent and correct punctuations (Lindvall and Nitco, 1975). Unfortunately, the test maker makes many mistakes in punctuations. According to Wood et al. (1991: 31), the use of proper grammar and punctuation, and spelling reduces ambiguity in the item and encourages students to take the test more seriously. They often missed the full stop (.) at the end of a sentence both in the stem and in the alternatives. The test makers should have paid more attention on it because twenty-one items have this problem.

Another punctuation that is misused is the comma. Lindvall and Nitco (1975) mention that punctuations must be used consistently. One type of items is not in line with this. More than one item is written in the form of rearranging the sentences provided in the stem. For the alternatives, the number of code for each of the sentences is written and they are separated with a dash but there is one item that is not. It means the test maker does not have consistency in the use of punctuation.

The other mistakes lie in the capitalization of the first letter of the options. According to Gronlund and Waugh (2009:106), when the stem is an incomplete statement, start each alternative with a lower case letter and end with whatever terminal punctuation mark is appropriate. There is also a grammatical mistake in the construction of interrogative sentences. This is not in line with Gronlund and Waugh’ statement (2009) that the test writer should follow the normal rules of grammar.

The second principle is providing clear directions for the test. Six of fifty items of the test have unclear directions. The next mistake was that the test makers did not follow the rule that there should not be any page change in the middle of an item. Wood et al. (1991: 30) mention that test makers must layout the items in a clear and consistent manner. They elaborate that good layout means providing clear directions at the beginning of each section of the test, avoid changing pages in the middle of an item and, in addition, using vertical format for presenting alternatives. The mistake will put students at disadvantage because items which do not follow both principles will make the test confusing and time consuming for students. It is obvious that for the items are
not in line with both principles in the guidelines.

The test maker also did not follow the fourth principle that an item must assess only one formulated problem. Both Gronlund and Waugh (2009) and Wood et al. (1991) disagree to put more than one problem in an item because that would add to the complexity of the wording and reduces the diagnostic value of the item (Gronlund and Waugh, 2009: 94). Moreover, items that are not written with a specific objective in mind often end up measuring lower-level objectives exclusively, or covering trivial material that is a little educational worth (Wood et al., 1991:15). Hence, the items are not in line with the guidelines.

Next, the test maker did not follow the fifth principle either. One item is confusing because the stem repeats the text stated previously. However, the content is a little different from the text. Students would be confused to infer which problems should be answered. This is not in line with Gronlund and Waugh’ idea (2009) that the stem of the item should be stated in simple and presented in clear language. According to Wood et al. (1991), any excess material in the stem that is not essential to answering the problem increases the reading burden and adds to student confusion over what he or she is being asked to do. Moreover, if the students have to infer what the problem is, the item will likely measure the students’ ability to draw inferences from vague descriptions rather than his or her achievement on the intended objective (Wood et al., 1991:15). Gronlund and Waugh (2009: 94-95) also state that it would increase reading time. Time spent in reading such irrelevant material could be spent more profitably in thinking about the problem presented.

The next principle is the stem should include words that otherwise would have to be repeated in each alternative. There are items with words repeated in each option. Rather than repeating words or phrases in each of the options, placing them in the stem will decrease the reading burden and more clearly define the problem in the stem (Wood et al., 1991: 16). Gronlund and Waugh (2009) also suggest that by moving all the common content to the stem, it is usually possible to clarify the problem further and to reduce the time the student needs to read the alternatives. They ensured that items function better when slim and trim. In addition, Lindvall and Nitko (1975) state that the responses should be made as short as possible.

The mistakes also occur in the eighth principle where the content of the alternatives must be homogeneous. This is important because if the alternatives consist of a potpourri of statements related to the stem but unrelated to each other, the
students’ task becomes unnecessarily confusing. Alternatives that are parallel in content help the item present a clear-cut problem and are more capable of measuring the attainment of a specific objective (Wood, et al., 1991: 19).

The last mistake is on the use of length of the alternatives as a clue. Usually, the longest alternative is the answer because of the need to qualify statements to make them unequivocally correct. Four items have the answer as the longest alternative. Lindvall and Nitko (1975) suggest that all the responses be of approximately the same length because if they were not, they would provide the examinee with an irrelevant clue to the correct answer.

CONCLUSIONS AND SUGGESTIONS

Conclusions

Based on the findings and discussions presented in the previous chapter, it can be concluded that the test was well developed based on the stages of test development. The stages were well-planned in the initial stages and the development process follows each stage completely. However, the stages they took are not exactly in line with the guidelines because they made a different variation of timeline of the stages. Some activities in the guideline were not carried out for the sake of practicality, such as trying out the test before it was used, gathering information of test takers’ test results, and analyzing the tests after it was used, and item banking.

The design of the scoring method was not well-planned. The test development team did not pay much attention to the scoring method because they limited their responsibility up to the test assembling. Since the test development team’s responsibility does not include the scoring of the test, they did not provide a rater team and the English teachers carried out the scoring activity. The design is appropriate with the type of test that is multiple-choice questions, especially in terms of practicality.

Suggestions

Based on the findings and discussions, the test is, in general, not well-constructed. Based on the analysis, there are some suggestions that can be proposed. It is suggested that teachers have to learn how to construct a good test by reading books about testing or joining workshops about testing and evaluation. Trying-out the test before it is used is important to know how well the test performs. If this is too
impractical to do, the researcher suggests that an analysis be done on the result of the
test after the test is held. This can be considered as the try-out for the next test. The
test development team must have an item bank to save all their works. The
availability of scorers is necessary to make sure that the students’ score are from an
objective marking. The scores can also provide information for the Department of
Education in Kabupaten Pasuruan about the students’ achievement. For further
researchers, it is suggested that they analyze the validity, reliability, and practicality
of the test made by MGMP Bahasa Inggris Kabupaten Pasuruan because they never
analyze the test they made in this way before.

REFERENCES

Alderson, J.C., Clapham C., and Wall, D. 1995. Language Test Construction and

Development and Examining. Strasbourg: Council of Europe, (Online),
downloaded on February 26, 2011.

Oxford University Press.

Bachman, L.F. and Palmer, A.S. 1996. Language Testing in Practice: Designing and


Applied Measurement in Education, 2(1), 79-96, (Online),


Lampiran Peraturan Menteri Pendidikan Nasional Nomor 20 tahun 2007 Tanggal
11 tahun 2007 tentang Standar Penerima Pendidikan, (Online),
(http://akhmadsudrajat.files.wordpress.com/), accessed on February 28,
2012.

Malang: English Department, State University of Malang.

Lindvall, C.M. and Nitko, A.J. 1975. Measuring Pupil Achievement and Aptitude


New York: Harper Collins Publisher.


