CHAPTER II

REVIEW OF TEST, RELIABILITY AND VALIDITY

This chapter presents the study review of test, reliability and validity. The chapter begins with the definition of test, the kind of test, the function of test, reliability, validity, item analysis, multiple choice items, curriculum and syllabus.

A. The Definition of Test

Test is very closely interrelated to the TLP. It is because test cannot be conducted without conducting the Teaching Learning Process before. While the success of the Teaching Learning Process cannot be measured without conducting the test. It means that test is an instrument or procedure to know or to measure something with the determined criteria. If it is related to the TLP, something that is measured is students' ability.

The function indicated in the preceding paragraph provides one of the answers to the question: Why test? But it must be emphasized that it is only one of the function of a test and that furthermore, as far as the practicing teacher is concerned, it is perhaps one of the more negative functions. Although most teacher also wish to evaluate individual performance, the aim of the classroom test is different to that of the external examination. While the latter is generally concerned with evaluation for the purpose of selection, the classroom test is concerned with evaluation for the purpose of enabling the teacher to increase his own effectiveness by making adjustments in his teaching to enable certain groups of students or individual in the class to benefit more. Too many teachers gear their teaching towards an ill-defined average group without taking into

account the abilities of those students in the class who are at either end of the scale.[1]

So, test is very important either for the teachers or for the students. The importance for the students through a test, they will know how far their achievement in learning the material. While for the teachers, through a test, they will know which students who have understood the material so that the teachers can give more attention to the students who have not understood yet. Test is any series of questions or exercises or means of measuring the skill, knowledge, intelligence, capacities of aptitudes or an individual or group. Test is comprehensive assessment of an individual or to an entire program evaluation effort.

B. The Kinds of Test

There are many kinds of tests. Based on Arikunto, kinds of test are viewed from its use in measuring the students' ability. There are 3 kinds of test: (1) Diagnostic test, (2) Formative test, and (3) Summative test. [2] The descriptions of each of them are as follows:

1. Diagnostic Test

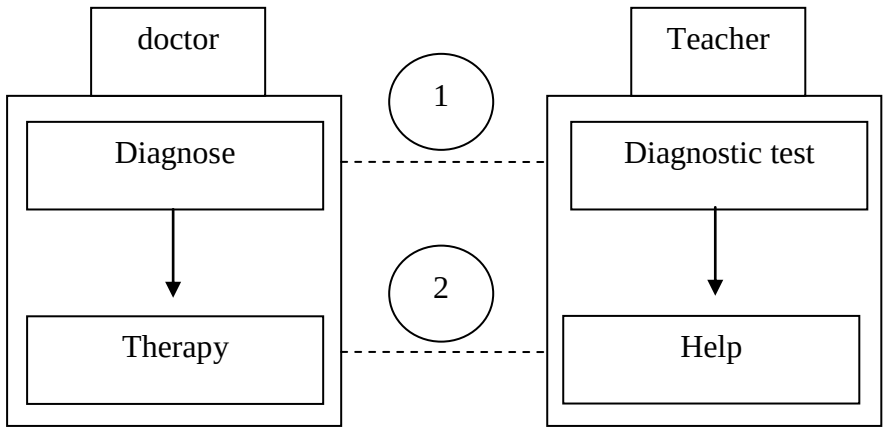
Diagnostic test is test used to know the weaknesses of the students in learning. So by knowing their weaknesses, the teacher can give the appropriate treatment. Besides it is important for the teacher in knowing the weaknesses of the students, through the diagnostic test, the students can also

[1] J. B. heaton, *Writing English Language Tests*, London: Longman Group Limited, 1975, p. 2
[2] Suharsimi Arikunto, *Dasar-dasar Evaluasi Pendidikan*, Jakarta: Bumi Aksara, 2009, p. 33.

know and be conscious with their weaknesses themselves so they can improve themselves better.[3]

Although the term diagnostic test is widely used, few tests are constructed solely as diagnostic tests. Achievement and proficiency tests, however, are frequently used for diagnostic purposes: areas of difficulty are diagnosed in such tests so that appropriate remedial action can be taken later.[4]

Example diagnostic test :



2. Formative Test

Formative test is a test that is conducted after one unit or one lesson finished given by the teacher. By conducting formative test, the teachers can know how far the success of the Teaching Learning Process especially for one lesson. So they can decide what the next actions to the students are.[5]

---

[3] *Ibid.*, p. 34.

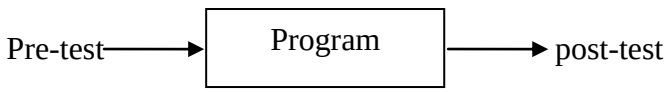
[4] J. B. heaton, *Writing English Language Tests*, London: Longman Group Limited, 1975, p. 165

[5] Suharsimi Arikunto, *Dasar-dasar Evaluasi Pendidikan*, Jakarta: Bumi Aksara, 2009, p. 36.

The next actions can be a treatment, a motivation, praise, etc., depending on the result got by the students in taking the formative test. Result here is score. If the score is good, the teacher can give praise so they will be motivated to get good score in the next lesson. While if the score is bad, the teacher can give more treatment especially in what the students' weaknesses, and also do not forget give motivation to the students to study better so they can get good score like their friends who have got good score. According to the statement above, the formative test can be diagnostic test. It is because besides to take score, conducting the formative test also to know the weaknesses of the students in learning material and the weaknesses of the teachers in giving material.

Formative test is generally carried out throughout a course or project. It is used to aid learning in that it helps the student and teacher to find out what the student knows so that the teacher can address any areas of weakness or misconceptions in subsequent lessons. The purpose of formative assessment is to see if students have mastered a given concept and can typically be assigned a pass/fail grade (if used for grading purposes at all).

Example formative test :



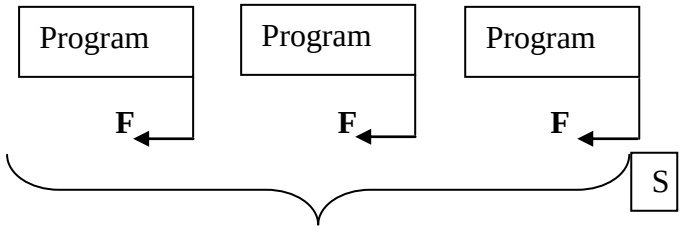
3. Summative Test

Summative test is a test that is conducted after all units are finished given by the teacher. This kind of test is conducted in the end of the

semester. The summative test is used to get educational decision. Educational decision means the students can pass or fail in mastering the material. The summative test is interrelated to the formative test because the subjects in the summative test are including all units or lessons which are tested in the formative test.[6]

Summative test is generally carried out at the end of a course or project. In an educational setting, summative are typically used to assign students a course grade, and often a scaled grading system enabling the teacher to differentiate students will be used.

Example summative test :



Note : F= Formative test S= Summative test

4. Discrete Language Test

Discrete Point tests are constructed on the assumption that language can be divided into its components parts, and those parts can be tested successfully. The components are the skills of listening, speaking, reading, writing, and various unit of language of phonology, morphology, lexicon, and syntax. Discrete point tests aim to achieve a high reliability factor by

[6] *Ibid*., p. 38.

testing a large number of discrete items, but each question tests only one linguistic point.[7]

According to Lado, discrete-point testing assumes that language knowledge can be divided into a number of independent facts: elements of grammar, vocabulary, spelling and punctuation, pronunciation, intonation and stress. These can be tested by pure items (usually multiple-choice recognition tasks). Integrative testing argues that any realistic language use requires the coordination of many kinds of knowledge in one linguistic event, and so uses items which combine those kinds of knowledge, like comprehension tasks, dictation, speaking and listening. Discrete-point testing risks ignoring the systematic relationship between language elements; integrative testing risks ignoring accuracy of linguistic detail.

5. Integrative Language Test

This approach involves the testing of language in context and is thus concerned primarily with meaning and the total communicative effect of discourse. Consequently, integrative tests do not seek to separate language skills into neat divisions in order to improve test reliability: instead, they are often designed to assess the learner's ability to use two or more skills simultaneously. Thus, integrative tests are concerned with a global view of proficiency - an underlying language competence or `grammar of expectancy", which it is argued every learner possesses regardless of the purpose for which the language is being learnt. Integrative testing involves

[7] Ratna KomalaDewi, approach in language testing,thejoyoflanguageassessment.wordpress .com.approaches-in-language-testing-2/ (online 08 Sept 2014)

`functional language'2 but not the use of functional language. Integrative tests are best characteristic by the use of doze testing and of dictation. Oral interviews, translation and essay writing are also included in many integrative tests - a point frequently overlooked by those who take too narrow a view of integrative testing.

There are two methods of scoring a doze test: one mark may be awarded for each acceptable answer or else one mark may be awarded for each er.icr answer. Both methods have been found reliable: some argue that the former method is very little better than the latter and does not really justify the additional work entailed in defining what constitutes an acceptable answer for each item. Nevertheless, it appears a fairer test for the student if any reasonable equivalent is accepted. In addition, no student should be penalized for misspellings unless a word is so badly spelt that it cannot be understood. [8]

6. Pragmatic Language Test

Pragmatic language ability involves the ability to appropriately use language (e.g., persuade, request, inform, reject), change language (e.g., talk differently to different audiences, provide background information to unfamiliar listeners, speak differently in different settings, etc) as well as follow conversational rules (e.g., take turns, introduce topics, rephrase sentences, maintain appropriate physical distance during conversational exchanges, use facial expressions and eye contact, etc) all of which

[8] J.B. Heaton, *Writing English Language Tests,* New York: Longman Group UK Limited, 1988, p.16.

culminate into the child's general ability to appropriately interact with others in a variety of settings.

For most typically developing children, the above comes naturally. However, for children with pragmatic language impairment appropriate social interactions are not easy. Children with pragmatic language impairment often misinterpret social cues, make inappropriate or off-topic comments during conversations, tell stories in a disorganized way, have trouble socially interacting with peers, have difficulty making and keeping friends, have difficulty understanding why they are being rejected by peers, and are at increased risk for bullying.

7. Communicative Language test

Communicative language test has recently introduced another dimension to the whole concept of the reliability: namely, profile reporting. In order to obtain a full profile of a student's ability in the target language, it is necessary to assess his or her performance separately for each of the different areas of communication: e.g. listening comprehension, speaking and listening, reading, reading and writing. (summaries, etc) and writing. Furthermore, performances are assessed according to t he purpose for which the language is to be used: e.g. academic, occupational, social survival. The object of the sub-tests through which performance is assessed is to indicate the extent of the learner's mastery of the various language skill which he or she will require for a particular purpose. A score or grade is given for each of the skills or areas selected for testing, and an average mark is eventually

obtained. This latter mark, however, is only given alongside the various scores which have contributed to it. Profile reporting is thus very valuable for placement purposes, and indeed it is an essential feature of one of the most widely used proficiency tests set in Britain and administered in many countries throughout the world. A student's performance on the various parts of a simple table or chart, in which the target score, appears beside the students' score. It is thus very easy to compare a student's performances levels in each area with the required levels.

The communicative approach to language testing is sometimes linked the integrative approach. However, although both approaches emphasise the importance of the meaning of utterances rather than their form and there are nevertheless fundamental differences between the two. Communicative tests are concerned primarily (if not totally) with how language is used in communication. Consequently, most aim to incorporate tasks which approximate as closely as possible to those facing the students in real life. Success is judged in terms of the effectiveness of the communication which takes place rather than formal linguistic accuracy.[9]

C. The Function of Test

In line with the Teaching Learning Process, simply, according to Sudijono, there are two functions of test. They are: (1) As an instrument to measure the development or progress that has been reached by the students after they go through the Teaching Learning Process within certain. (2) As an

[9] J.B. Heaton, *Writing English Language Tests*, New York: Longman Group UK Limited, 1988, p.19.

instrument to measure the success of the Teaching Learning Process. Through the test, the teachers will be able to know how far the programmed materials have been able to be reached by the students. [10]

A good classroom test will also help to locate the practice area of difficulty encountered by the class or by the individual students. Just as it equally necessary for the doctor first to diagnose his patient's illness, so it is equally necessary for the teacher to diagnose his student's weaknesses and difficulties. Unless the teacher is able to identify and analysis the errors a student makes in handling the target language, he will be in no position to render any assistance at all through appropriate anticipation, remedial work and additional practice. The test should also enable the teacher to ascertain which parts of the language program have been found difficult by the class. In this way, the teacher can evaluate the effectiveness of the syllabus as well as the methods and materials he is using. The test result may indicate, for example, certain areas of the language syllabus which have not taken sufficient account of LI learner difficulties or which, for some reason, have been glossed over.

A test which sets out to measure a student's performance as fairly as possible without in any ways setting traps for him can be effectively used to motivate the student. A well-constructed classroom test will provide the students. With an opportunity to show his ability to recognize and produce correct forms of the language. Provided that details of his performance area given as soon as possible after the test, the student should be able to learn from

[10] Anas Sudijono, *Pengantar Evaluasi Pendidikan*, Jakarta, RajaGrafindo Persada, 2011, p. 67.

his errors and consolidate the pattern taught. In this way a good test can be used as a valuable teaching devise.[11] So, the function of the test is important either for the teachers or for the students. The former is important for the students and the latter is important for the teacher.

D. Reliability

1. Definition of Reliability

J. B. Heaton stated "Reliability is a necessary characteristic of any good test. For it to be valid at all, a test must first be reliable as a measuring instrument". If the test is administered to the same candidates on different occasions (with the extent that it produces differing results, it is not reliable. Reliability measured in this way is commonly referred to as test/re-test reliability to distinguish it from mark/re-mark reliability and the other kind of reliability denotes the extent to which the same marks or grades are awarded if the same test papers are marked by two or more different examiners or the same examiner on different occasions. In short, in order to be reliable a test must be consistent in its measurements.

Reliability is of primary importance in the use of both public achievement and proficiency tests and classroom tests. However, an appreciation of the various factors affecting reliability is important for the teacher at the very outset, since many teachers tent to regard tests as infallible measuring instruments and fail to realize that even the best test is indeed an imprecise instrument with which to measure language skills. Reliability refers

[11] J. B. heaton, *Writing English Language Tests,* London: Longman Group Limited, 1975, p. 2

to the consistency of the scores obtained-how consistent they are for each individual from one administration of an instrument to another and from one set of items to another. Consider, for example, a test designed to measure typing ability. If the test is reliable, we would expect a student who receives a high score the first the first time he takes the test to receive a high score the next time he takes the test. The scores would probably not be identical, but they should be close.

The score obtained from an instrument can be quite reliable but not valid. Suppose a researches gave a group of eleventh-graders two forms of a test designed to measure their knowledge of the constitution of the SMA Muhammadiyah 1 Palangka Raya and found their score to be consistent: those who scored high on form A also scored high on form B; those who scored low on A scored low on B; and so on. We would say that the score were reliable. But if the researcher then used these same test scores to predict the success of these students in their physical education classes, she would probably be looked at in amazement.

Factors affecting the reliability of a test are :

1) The extent of the sample of material selected for testing: whereas validity is concerned chiefly with the content of the sample, reliability is concerned with the size. The large the sample (i.e. the more tasks the testee has to perform), the greater the probability that the test as a whole is reliable-hence the favoring of objective test, which allow for a wide field to be covered.

2) The administration of the test: is the same test administered to different groups under different conditions or at different times? Clearly, this is an important factor in deciding reliability. Especially in tests of oral production and auditory comprehension.

   The way in which this factor differ from test situation validity can be seen from the following example : if a recording for an auditory comprehension test is initially poor in quality, then it is poor in quality for all testee. This will consequently make for invalidity fun less speech has been deliberately masked with noise as a testing device. But if the quality of the recording is good and if certain group and if certain groups hear it played under good acoustic conditions while other groups hear it under poor acoustic condition, this will make for unreliable and therefore invalidity.

3) Test instruction: are the various tasks expected from the testee made clear to all candidates in the public?

4) Personal factors such as motivation and illness.

5) Scoring the test: one of the most important factors affecting reliability. Objective tests overcome this problem of marker reliability, but subjective tests are still faced with it: hence the importance of the work carried out in the field of the multiple-marking of composition.

One method of measuring the reliability of a test is to re-administer the same test after a lapse of time. It is assumed that all candidates have been treated in the same way in the interval-that they have either all been taught or

that none of them have. Provided that such as assumptions (which are frequently hard to justify) can be made, comparison of the two result would then show how reliable the test has proved. Clearly, this method is often impracticable and, in any case, a frequent use of it is not to be recommended, since certain students will benefit more than others by a familiarity with the type and formal of the test. Moreover, in addition to changes in performance resulting from the memory factor, personal factors such as motivation and differential maturation will also account for differences in the performances of certain students.

Another means of estimating the reliability of a test is by administering parallel forms of the test to the same group. This assumes that two similar version of a particular teat can be constructed: such as test must be identical in the nature of their sampling, difficulty, length, rubrics, etc. only after a full statistical analysis of the test and all the items contained in them can the tests safely be regarded as parallel. If the correlation between the two test is high , then the test can be termed reliable.

Reliability refers to the consistency of a measure. A test is considered reliable if we get the same result repeatedly. For example, if a test is designed to measure English ability, the results should be approximately the same. Unfortunately, it is impossible to calculate reliability exactly, but there several different ways to estimate reliability.

In line with the statements above, Baron and Bernard stated, "The consistency with which a test measures whatever it does measure is called reliability"[12].

So, based on some experts above, if it is related to the test, especially the English summative test, that reliability is the consistency of the score in whenever the English summative test is conducted.

2. Character of Good Reliability

A. Equivalent method (Parallel Form)

In parallel forms reliability you first have to create two parallel forms. One way to accomplish this is to create a large set of questions that address the same construct and then randomly divide the questions into two sets. You administer both instruments to the same sample of people. The correlation between the two parallel forms is the estimate of reliability. One major problem with this approach is that you have to be able to generate lots of items that reflect the same construct. This is often no easy feat. Furthermore, this approach makes the assumption that the randomly divided halves are parallel or equivalent. Even by chance this will sometimes not be the case. The parallel forms approach is very similar to the split-half reliability described below. The major difference is that parallel forms are constructed so that the two forms can be used independent of each other and considered equivalent measures.

[12] Denis Baron and Harold W. Bernard, *Evaluation Techniques for Classroom Teachers*, New York, McGraw-Hill Book Company, Inc., 1958, p. 20.

B. Test-retest Method

Estimate test-retest reliability when we administer the same test to the same sample on two different occasions. This approach assumes that there is no substantial change in the construct being measured between the two occasions. The amount of time allowed between measures is critical. We know that if we measure the same thing twice that the correlation between the two observations will depend in part by how much time elapses between the two measurement occasions. The shorter the time gap, the higher the correlation; the longer the time gap, the lower the correlation. This is because the two observations are related over time -- the closer in time get the more similar the factors that contribute to error. Since this correlation is the test-retest estimate of reliability, you can obtain considerably different estimates depending on the interval.

C. Split-half Method

The split-half method is yet another means of measuring test reliability. This method estimates a different kind of reliability from that estimated by test/re-test procedures. The split-half method is based on the principle that, if an accurate measuring instrument were broken into two equal parts, the measurements obtained with the other. The test is divided into two and the corresponding scores obtained, the extent to which they

correlate with each other governing the reliability of the test as a whole. [13]

Calculating the reliability of the test by using the Kuder-Richardson's formula (KR-20).

$$r_{11} = \left(\frac{n}{n-1}\right)\left(\frac{S^2 - \Sigma pq}{S^2}\right)$$

Where:

r_{11} : Instrument reliability
n : The number of items in the test
S : Total variance
p : Proportion of correct answer
q : Proportion of wrong answer (q=1-p)
Σpq : summary of p multiply q

The interpretations of reliability coefficient based on Sudijono are as follows:

≥ 0.70 : reliable
< 0.70 : unreliable[14]

E. Validity

According to Cronbach," how well a test or evaluative technique does the job that it is employed to do".[15] Briefly, the validity of a test is the extent to which it measures what it is supposed to measure and nothing else. Every test whether it be a short, informal classroom test or a public examination. Should be as valid as the constructor can make it. The test must aim to provide a true measure of the particular skill which it is intended to measure: to the extent that

[13] J. B. heaton, *Writing English Language Tests*, London: Longman Group Limited, 1975, p. 156

[14] Suharsimi, *Dasar-dasar Evaluasi Pendidikan*, Jakarta: Bumi Aksara, 1995, p. 98

[15] Ngalim Purwanto, *Prinsip-prinsip Tekjnik Evaluasi*, Bandung; Ramadja Karya, 1984, p. 117

it measure external knowledge and other skills at the same time, it will not be a valid test.

Based on J. B. Heaton "the validity of a test is the extent to which it measure what it is supposed to measure nothing else". A test that is given should be valid. Valid test means the test can really measure what it is intended to measure, not else. For example, if the teacher would like to measure speaking ability of the students, so the teacher should give question orally and the students should also answer orally.[16]

In line with the statements above, Tuckman stated, "Test validity refers to whether a test measures what we intend it to measure". [17] In the case of this study, it is the English summative test, it should be valid. The English summative test should really measure what it is intended to measure. Validity is the most important idea to consider when preparing or selecting an instrument for use. More than anything else, researchers want the information they obtain through the use of an instrument to serve their purposes.

In recent years, validity has been defined as referring to the appropriateness, correctness, meaningfulness, and usefulness of the specific inferences researches make based on the data they collect. Validation is the process of collecting and analyzing evidences to support such inferences. The important point here is to realize that validity refers to the degree to which evidence supports any inferences a researches makes based on the data he or

---

[16] J. B. heaton, *Writing English Language Tests*, London: Longman Group Limited, 1975, p. 153

[17] Bruce W. Tuckman, *Measuring Educational Outcomes Fundamentals of Testing*, USA, Harcourt Brace Jovanovich, Inc., 1975, p. 229.

she collect using a particular instrument. It is the inferences about the specific uses of an instrument that are validated, not be instrument itself. These inferences should be appropriate, meaningful, correct, and useful.

One interpretation of this conceptualization of validity has been that test publishers no longer have a responsibility to provide evidence of validity. We do not agree; publishers have an obligation to state what an instrument is intended to measure and to provide and evidence that it does. Nonetheless, researchers must still give attention to the way in which they intend to interpret the information.[18] In this study, the writer will focus on logical validity and empirical validity. The kinds of validity and they are discussed as follows:

1. Logical Validity

1) Content Validity

According to J. B. Heaton "this kind of validity depends on a careful analysis of the language being tested and of the particular course objectives".[19] The test should be so constructed as to contain a representative sample of the course, the relationship between the test item and the course objective always being apparent. There is a strong tendency, especially in multiple-choice testing, to test only those areas of language which lend themselves readily to testing.

---

[18] Jack R. Fraenkel, *How To Design And Evaluate Research in Education,* t.tp., t.np., 2007, sixth edition, p. 150.

[19] J. B. heaton, *Writing English Language Tests*, London: Longman Group Limited, 1975, p. 154

A test fulfills the content validity if it measures the materials that have been programmed and given. The programmed materials are as described in the curriculum. So, to reach the content validity, it is necessary to construct the items of the test based on the materials that have been programmed in curriculum. The details of the curriculum can be seen in syllabus.

2) Construct Validity

Based on Riduan "if a test has construct validity, it is capable of measuring certain specific characteristic in accordance with a theory of language behavior and learning". Construct validity refers to the practical tests developed from a theory[20]: Based on J. B. Heaton "if a test has construct validity, it is capable of measuring certain specific characteristic in accordance with a theory of language behavior and learning". [21]

A test fulfills the construct validity if the items of the test constructed to measure every aspect of learning such as mentioned in the Specific Instructional Objectives *(Tujuan Instruksional Khusus).* Specific Instructional Objectives as Arikunto stated above, in School-Based Curriculum or *Kurikulum Tingkat Satuan Pendidikan (KTSP)* are equivalent to the Indicators in the Lesson Plan or curriculum or syllabus that show the success of the students in mastering the material it can be concluded that the content validity relates to the material in the

[20] Riduan, *Metode dan teknik Menyusun Thesis,*Bandung: Alfabetap, 2004, p. 110.
[21] *Ibid.*, p. 154.

curriculum or the syllabus and the construct validity relates to the indicators in the curriculum or the syllabus. If it relates to this study, the English summative test ideally have fulfilled content validity and construct validity.

2. Empirical Validity

A fourth type of validity is usually referred to as statistical or empirical or empirical validity. This validity is obtained as a result of comparing the result of the test with the result of some criterion measure such as :

a. An existing test, known or believed to be valid and given at the same time; or
b. The teacher's ratings or any other such form of independent assessment given at the same time, or
c. The subsequent performance of the testes on a certain task measured by some valid test, or
d. The teacher's rating or any other such form of independent assessment given later.

Results obtained by either of the first two methods above are measure of the test's concurrent validity in respect of the particular criterion used. The third and fourth method estimates the predictive validity of the test which is used to predict future success. The test situation or the technique used is always an important factor in determining the overall validity of any test. Although an ideal test situation will by no means guarantee validity, a poor

test situation will certainly detract from it. Is an auditory comprehension test valid if the testee hears only disembodied voice on, say, a poor quality tape-record?[22]

There are empirical validity:

1) Concurrent Validity

Concurrent validity occurs when the criterion measures are obtained at the same time as the test scores. This indicates the extent to which the test scores accurately estimate an individual's current state with regards to the criterion. For example, on a test that measures levels of depression, the test would be said to have concurrent validity if it measured the current levels of depression experienced by the test taker. In concurrent validity, assess the operationalization's *ability to distinguish between groups that it should theoretically be able to distinguish between*. For example, if we come up with a way of assessing manic-depression, our measure should be able to distinguish between people who are diagnosed manic-depression and those diagnosed paranoid schizophrenic. If we want to assess the concurrent validity of a new measure of empowerment, might give the measure to both migrant farm workers and to the farm owners, theorizing that our measure should show that the farm owners are higher in empowerment. As in any discriminating test, the results are more powerful if you are able to show that you can discriminate between two groups that are very similar.

[22] *Ibid.,* p. 155.

2) Predictive Validity

Predictive Validity occurs when the criterion measures are obtained at a time after the test. Examples of test with predictive validity are career or aptitude tests, which are helpful in determining who is likely to succeed or fail in certain. In predictive validity, we assess the operationalization *ability to predict something it should theoretically be able to predict*. For instance, we might theorize that a measure of math ability should be able to predict how well a person will do in an engineering-based profession. We could give our measure to experienced engineers and see if there is a high correlation between scores on the measure and their salaries as engineers. A high correlation would provide evidence for predictive validity -- it would show that our measure can correctly predict something that we theoretically think it should be able to predict.

3. Validity Item

According to Sugiono "correlation techniques to determine the validity of this item until the most widely used technique".[23] The validity of a test is the extent to which it measures what it is supposed to measure and nothing else. A test that is given should be valid. Valid test means the test can really measure what it is intended to measure, not else. To measuring validity item there are two formulas by Pearson.

[23] Sugiyono, *Metode Penelitian Pendidikan,* Bandung: Alfabeta, 2007, p. 188.

a. product moment correlation with deviation

$$r_{xy} = \frac{\Sigma xy}{\sqrt{(\Sigma x^2)(\Sigma y^2)}}$$

Where:

r_xy	=	Coefficient correlation between variable X and variable Y. Two variable correlation (x=X-M) and (y= Y-M).
∑xy	=	summary of x multiply y
x^2	=	square x (deviation x)
y^2	=	square y (deviation y).[24]

b. product moment correlation with a rough figure

$$r_{xy} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{(N\Sigma X^2-(\Sigma X)^2)(N\Sigma Y^2-(\Sigma Y)^2)}}$$

Where:

r_xy	=	Coefficient correlation between variable X and variable Y
Σxy	=	summary of x multiply y
∑x^2	=	total square x
∑y^2	=	total square y
(∑x)^2	=	total X then square
(∑y)^2	=	total Y then square.[25]

positive correlation if *x* and *y* have a strong positive linear correlation, *r* is close to +1. An *r* value of exactly +1 indicates a perfect positive fit. Positive values indicate a relationship between *x* and *y* variables such that as values for *x* increases, values for *y* also increase. Negative correlation if *x* and *y* have a strong negative linear correlation, *r* is close to -1. An *r* value of exactly -1 indicates a perfect negative fit. Negative values indicate a relationship between *x* and *y* such that as values for *x* increase, values for *y* decrease. Coefficient correlation:

---

[24] Suharsimi, *Dasar-dasar Evaluasi Pendidikan*, Jakarta: Bumi Aksara, 1995, p. 67

[25] *Ibid.* p. 69

- ♥ between 0,800 to 1,00 : very high
- ♥ between 0,600 to 0,800 : high
- ♥ between 0,400 to 0,600 : medium
- ♥ between 0,200 to 0,400 : low
- ♥ between 0,00 to 0,200 : very low

F. Item Analysis

Earlier careful consideration of objectives and the compilation of a table of test specifications were urged before the construction of any test was attempted. What is required now is knowledge of how far those objectives have been achieved by a particular test. Unfortunately, too many teachers think that the test is finished once the raw marks have been obtained. But this is far from the case, for the results obtained from objective tests can be used to provide valuable information concerning:

- ♥ the performance of the students as a group, thus (in the case of class progress tests) informing the teacher about the effectiveness of the teaching;
- ♥ the performance of individual students; and
- ♥ the performance of each of the items comprising the test.

Information concerning the performance of the students as a whole and of individual students is very important for teaching purposes, especially as many test results can show not only the types of errors most frequently made but also the actual reasons for the errors being made. As shown in earlier chapters, the great merit of objective tests arises from the fact that they can provide an

insight into the mental processes of the students by showing very clearly what choices have been made, thereby indicating definite lines on which remedial work can be given.

The performance of the test items, themselves, is of obvious importance in compiling future tests. Since a great deal of time and effort are usually spent on the construction of good objective items, most teachers and test constructors will be desirous of either using them again without further changes or else adapting them for future use. It is thus useful to identify those items which were answered correctly by the more able students taking the test and badly by the less able students. The identification of certain difficult items in the test, together with knowledge of the performance of the individual distracters in multiple-choice items, can prove just as valuable in its implications for teaching as for testing.

All items should be examined from the point of view of (1) their level of difficulty (2) their level of discrimination (3) function distracters.

1. Level of Difficulty

   Based on Crocker and Algina "level of difficult is correct answer from the student take the test". According to Mulyasa, "test will be affective if you use that approach the level of difficulty on the acceptance (cut of score)".[26] According to Sumadi Suryabrata, more educate level of difficulty is a price index of difficult is the transformation Z, the proportion of correct answer to the raw score. Based on J. B. Heaton "the index of difficulty of an

[26] Mulyasa, *Analisi Validitas, Reliabilitas dan Interprestasi Hasil test,* Bandung; RemajaRosdakarya, p. 22

item simply shows how easy or difficult the particular item proved in the test". The index of difficulty (or facility value) of an item simply shows how easyor difficult the particular item proved in the test. The index of difficulty (P) is generally expressed as the fraction (or percentage) of the students who answered the item correctly. It is calculated by using the formula:

$$P = \frac{B}{JS}$$

Where :

P = level of difficulty

B = represents the number of correct answers

JS = the number of students taking the test.[27]

Difficulty index :

0,0 ——————— 1,0

Difficult easy

- ♥ P 1,00 to 0,30 difficult
- ♥ P 0,30 to 0,70 medium
- ♥ 0,70 to 1,00 easy

2. Level of Discrimination

According to Mulyasa, "first stage to calculate the level of discrimination is to determine the upper and lower groups. Generally the test experts divide this group into 27% or 33% upper group and 27% or 33%

[27] Suharsimi, *Dasar-dasar Evaluasi Pendidikan*, Jakarta: Bumi Aksara, 1995, p. 214

lower group".[28] The discrimination index of an item indicates the extent to which the item discriminates between the testees, separating the more able testees from the less able. The index of discrimination (D) tells us whether those students who performed well on the whole test tended to do well or badly on each item in the test. It is presupposed that the total score on the test is a valid measure of the student's ability (i.e the good student tends to do well on the test as a whole and the poor student badly). On this basis, the score on the whole test is accepted as the criterion measure, and it thus becomes possible to separate the 'good' students from the 'bad' ones in performances on individual items. If the 'good' students tend to do well on an item (as shown by many of them doing so a frequency measure) and the `poor' students badly on the same item, then the item is a good one because it distinguishes the `good from the `bad' in the same way as the total test score. This is the argument underlying the index of discrimination.

There are various methods of obtaining the index of discrimination: all involve a comparison of those students who performed well on the whole test and those who performed poorly on the whole test. However, while it is statistically most efficient to compare the top 27 percent with the bottom 27 percent, it is enough for most purposes to divide small samples (e.g. class scores on a progress test) into halves or thirds. For most classroom purposes, the following procedure is recommended.[29]

---

[28] Mulyasa, *Analisi Validitas, Reliabilitas dan Interprestasi Hasil test,* Bandung; RemajaRosdakarya, p. 24.

[29] J. B. heaton, *Writing English Language Tests,* London: Longman Group Limited, 1975, p. 173.

Formula discrimination:

$$D = \frac{B_A}{J_A} - \frac{B_B}{J_B} = P_A - P_B$$

Where :

D	=	Discrimination
J	=	the number of students taking the test
J_A	=	total upper group
J_B	=	total lower group
B_A	=	total upper group the number of correct answer
B_B	=	total lower group the number of correct answer
P_A	=	proportion upper group the number of correct answer (P as difficulty index)

P_B = proportion lower group the number of correct answer

Classification discrimination:

- D : 0,00 to 0,20 : Poor
- D : 0,20 to 0,40 : satisfactory
- D : 0,40 to 0,70 : good
- D : 0,70 to 1,00 : excellent[30]

3. Analysis Distracters

It will often be important to scrutinize items in greater detail, particularly in those cases where items have not performed as expected. We shall want to know not only why these item have not performed according to expectation but also why certain testees have failed to answer a particular item correctly. Such task are reasonably simple and straight forward to perform if the multiple technique has been used in the test.

[30] Ibid., 218

In order to carry out a full item analysis, or an extended answer analysis, a record should be made of the different options chosen by each student in the upper group and then the various options selected by the lower group.

G. Multiple Choice Items

1. General

It is useful at this stage to consider multiple-choice items in some detail, as they are undoubtedly one of the most widely used types of items in objective tests. However, it must be emphasis at the outset that the usefulness of this type of item is limited. Unfortunately, multiple-choice testing has proliferated as a result of attempts to use multiple-choice items to perform tasks for high they were never intended Moreover, since the multiple-choice item is one of the most difficult and time-consuming types of items to construct, numerous poor multiple-choice tests now around. Indeed, the length of time required to construct good multiple-choice items could often have been better spent by teachers on other more useful tasks connected with teaching or testing.

The chief criticism of the multiple-choice item, however, is that frequently it does not lend itself to the testing of language as communication. The process involved in the actual selection of one out of four or five options bears little relation to the way language is used in most real-life situations. Appropriate responses to various stimuli in everyday situations are produced rather than chosen from several options.

Multiple choice items can provide useful means of teaching and testing in various learning situations (particularly at the lower levels) provided that it is always recognize that such items test of grammar, vocabulary, etc. rather than the ability to use language. Although they rarely measure communication as such, they can prove useful in measuring students' ability to recognize correct grammatical forms, etc. and to make important discriminations in the target language. In doing this, multiple-choice items can help both student and teacher to identify areas of difficulty.

Furthermore, multiple-choice items offer a useful introduction to the construction of objective tests. Only through an appreciation and mastery of the techniques of multiple-choice item writing is the would-be test constructor fully able to recognize the limitations imposed by such items and then employ other more appropriate techniques of testing for certain purposes.

The optimum number of alternatives, or options, for each multiple-choice item is five in most public tests. Although a larger number, say seven, would reduce even further the element of chance, it is extremely difficult and often impossible to construct as many as seven good options. Indeed, since it is often very difficult to construct items with even five options, four options are recommended for most classroom tests. Many writers recommend using four options for grammar items, but five for vocabulary and reading.

Before constructing any test items, the test writer must first determine the actual areas to be covered by multiple-choice items and the number of

items to be-included in the--test-The test must be long enough to allow for a reliable assessment of a testees performance and short enough to be practicable. Too long a test is undesirable because of the administration difficulties often created and because of the mental strain and tension which may be caused among the students taking the test. The number of items included in a test will vary according to the level of difficulty, the nature of the areas being tested, and the purpose of the test. The teacher's own experience will generally determine the length of a test for classroom use, while the length of a public test will be affected by various factors, not least of which will be its reliability measured statistically from the results of the trial test.

Note that context is of the utmost importance in all tests. De context multiple-choice items can do considerable harm by conveying the impression that language can be learnt and used free of any context. Both linguistic context and situational context are essential in using language. Isolated sentences in a multiple-choice test simply add to the artificiality of the test situation and give rise to ambiguity and confusion. An awareness of the use of language in an appropriate and meaningful so essential a part of any kind of communication then becomes irrelevant in the test. Consequently, it is important to remember that the following multiple-choice items are presented out of context here simply in order to save space and to draw attention to the salient points being made.

The initial part of each multiple-choice item is known as the stem; the choices from which the students select their answers are referred to as options/responses/alternatives. One option is the answer, correct option or key, whil6 the other options are distracters. The task of a distracter is to distract the majority of poor students (i.e. those who do not know the answer) from the correct option.

Stay here until Mr. Short ..... you to come. = stem

A. Told
B. Will tell
C. Is telling
D. Tells

Options

Distracter

= correct option/key

The following general principles should be observed when multiple choice items are constructed:

1) Each multiple-choice item should have only one answer. This answer must be absolutely correct unless the instruction specifies choosing the best option (as in some vocabulary tests). Although this may seem an easy matter, it is sometimes extremely difficult to construct an item having only one correct answer. An example of an item with two answers is:

'I stayed there until John ...

A. Had
B. Will
C. Would

D. Has

2) Only one feature at a time should be tested: it is usually less confusing for the testees and it helps to reinforce a particular teaching point. Obviously, few would wish to test both grammar and vocabulary at the same time, but sometimes word order and sequence of tenses are tested simultaneously. Such items are called impure items:

I never knew where…..

A. Had the boys gone
B. The boys have gone
C. Have the boys gone
D. The boy hade gone

3) Each option should be grammatically correct when placed in the stem, except of course in the case of specific grammar test items. For example, stems ending with the determiner a, followed by options in the form of nouns or noun phrases, sometimes trap the unwary test constructor. In the item below, the correct answer C, when moved up to complete the stem, makes the sentence grammatically incorrect:

Someone who designs houses is a…….

A. Designer
B. Builder
C. Architect
D. Plumber

The item can be easily re-cast as follow :

Someone who designs houses is a…….

A. A designer

B. A builder

C. An architect

D. A plumber

Steam ending in are, were, etc. may have the same weaknesses as the following and will require complete re-writing :

The boy hobbies referred to in the first paragraph of the passage were

A. Camping and fishing

B. Tennis and golf

C. Cycling long distances

D. Fishing, rowing and swimming

E. Collection stamps

Any fairly intelligent student would soon be aware that options C and E were obviously not in the tester's mind when first constructing the item above because they are ungrammatical answers. Such a student would, therefore, realize that they had been added later simply as distracters. Stems ending in prepositions may also create certain difficulties. In the following reading comprehension item, option C can be ruled out immediately:

John soon return to…….

A. Work

B. The prison

C. Home

D. School

4) All multiple-choice items should be at a level appropriate to the -proficiency level of the testees. The context, itself, should be at a lower level than the actual problem which the item is testing: a grammar test item should not contain other grammatical features as difficult as the area being tested, and a vocabulary item should not contain more difficult semantic features in the stem than the area being tested.

5) Multiple-choice items should be as brief and as clear as possible (though it is desirable to provide short contexts for grammar items).

6) In many tests, items are arranged in rough order of increasing difficulty. It is generally considered important to have one or two simple items to `lead in' the testees, especially if they are not too familiar with the kind of test being administered. Nevertheless, areas of language which are trivial and not worth testing. Should be excluded from the-test. idea of the problem and the answer required. At the same time, the stem should not contain extraneous information or irrelevant clues, thereby confusing the problem being tested. Unless students understand the problem being tested, there is no way of knowing whether or not they could have handled the problem correctly. Although the stem should be short, it should convey enough information to indicate the basis on which the

correct option should be selected.[31]

2. The Steam/The Correct Option/The Distracters

a. The Steam

I. The primary purpose of the steam is to present the problem clearly and concisely. The testee should be able to obtain from the steam a very general idea of the problem and the answer required. At the same time, the steam should not contain extraneous information or irrelevant clues, thereby confusing the problem being tested. Unless the student understands the problem being tested, there is no way of knowing whether or not he could have handed the problem correctly. Although the steam should be short, it should convey enough information to indicate the basic on which the correct option should be selected.

II. The stem may take the following forms:

(a) An incomplete statement

He accused me of ... lies.

A. speaking B. saying C. telling D. talking

(b) A complete statement

Everything we wanted was to hand.

A. under control C. well cared for

[31] J. B. heaton, *Writing English Language Tests*, London: Longman Group Limited, 1975, p. 16.

B. within reach D. being prepared

(c) A question

According to the writer, what did Tom immediately do?

A. He ran home. C. He began to shout.

B. He met Bob. D. He phoned the police.

III. The stem should usually contain those words or phrases which Mould otherwise have to he repeated in each option.

The word 'astronauts' is used in the passage to refer to…..

A. travelers in an ocean liner

B. travelers in a space-ship

C. travelers in a submarine

D. travelers in a balloon

The stern here should be rewritten so that it reads:

The word 'astronauts' is used in the passage to refer to travelers in

A. an ocean liner C. a submarine

B. a space-ship D. a balloon

The same principle applies to grammar items. The following item:

I enjoy ..... the children playing in the park.

A. looking to C. looking about

B. looking at D. looking on should be rewritten in this way:

I enjoy looking ..... the children playing in the park.

A. to B. about C. at D. on

If, however, one of the errors made by students in their free written work has been the concision of the preposition aft; r look (a common error), then it will be necessary to include look in the options·

I enjoy ..... the children playing in the park.

A. looking on C. looking at

B. looking D. looking to

IV. The stem should allow for the 'number of choices which have been decided upon. This is particularly relevant, for example, when comparisons are involved in reading comprehension. There is no possible fourth option which can be added in the following item:

Tom was ... the other two boys.

A. taller than B. smaller than C. as tall as

b. The correct option

For normal purposes of testing, this should be clearly the correct or best option: thus, it is most important that each item should be checked by another person. It can be argued that a greater degree of subtlety is sometimes gained by having more than one correct option in each item. The correct answers in the following reading comprehension and grammar items are circled:

According to the writer, Jane wanted a new racquet because

A. her old one was damaged slightly

B. she had lost her old one

C. her father had given her some money for one

D. Mary had a new racquet

E. Ann often borrowed her old racquet

Who…. You cycle here to see us.

A. ordered B. caused C. made D. asked E. let

It is very important, however, to avoid confusing the students by having a different number of correct options for each item, and this practice is not recommended. Each of the two multiple-choice test items above actually comprises a group of true/false (i.e. right/wrong) items and, therefore, each alternative should be marked in this way: e.g. in the first item, the testee scores I mark for circling A, 1 mark for not circling B, 1 mark for not circling C, I mark for circling D, and I mark for not circling E (total score = 5).

The correct option should be approximately the same length as

c. The distracters

Each distracter, or incorrect option, should be reasonably attractive and plausible. It should appear right to any testee who is unsure of the correct, option. Items should be constructed in such a way that students obtain the correct option-b--direct selection-rather than by the elimination of obviously incorrect options. Choice D in the following

grammar item is much below the level being tested and will be eliminated by testees immediately: their chances of selecting the correct option will then be one in three.

The present tax reforms have benefited …. poor.

A. that B. the C. a D. an

For most purposes, each distracter should be grammatically correct when it stands by itself: otherwise testees will be exposed to incorrect forms. In the above item (and in all grammar items) it is only the wrong choice, and its implied insertion into the stem, which makes a particular pattern ungrammatical. For example, option A is grammatically correct on its own and only becomes incorrect when inserted into the stem.

The following item (which actually appeared in a class progress test of reading comprehension) contains two absurd items:

How did Picard first travel in space?

A. He travelled in a space-ship.

B. He used a large balloon.

C. He went in a submarine.

D. He jumped from a tall building.

Unless a distracter is attractive to the student who is not sure of the correct answer, its inclusion in a test item is superfluous. Plausible distracters are best based on (a) mistakes in the students' own written work, (b) their answers in previous tests, (c) the teacher's experience, and

(d) a contrastive analysis between the native and target languages. Distracters should not he too difficult nor demand a higher proficiency in the language than the correct option. If they are too difficult, they will succeed only in distracting the good student. who will be led into considering the correct option too easy (and a trap). There is a tendency for this to happen, particularly in vocabulary test items.

You need a ..... to enter that military airfield.

A. permutation B. perdition C. permit D. perspicuity

H. Multiple Choice Items (Reading)

1. Short Texts

Type 1 It can be argued that the type of item in this section is in many ways a test of vocabulary rather than of reading comprehension. These particular items, however, have been included here because it is felt that a comprehension of the text is generally of at least as much importance as an understanding of the meaning of the words for selection. This, of course, is true of any vocabulary item presented in context: however, here the emphasis is more on the correct understanding of the context. The following three examples show the use of this item type at elementary, intermediate and advanced levels respectively.

1) The eyes are wonderful teachers - even musicians, who deal with sound, learn as much by (doing, playing, watching, practising) as by listening.
2) The housewife who could not afford to buy clothes would spend hours at her spinning wheel, spinning her wool into yarn a job which took little

skill but required a lot of (ability, patience, talent, wisdom) and was done by the fireside during the long winter evenings.

3) Two-thirds of the country's (fuel, endeavor, industry, energy) comes from imported oil, while the remaining one-third comes from coal. Moreover, soon the country will have its first nuclear power station.

Type 2 Just as the previous item type is closely related to the testing of vocabulary, so this type is perhaps more accurately described as a test of comprehension of grammatical structure. The testees are required to identify the correct paraphrase of a statement from a choice of four or five. They are told in the rubric that the (four) statements may refer to the entire sentence or only part of the sentence. Again, examples are provided for each of the three general levels.

1. John is not as tall as Sally but he's a little taller than Rick.

A. Sally is taller than John and Rick.

B. John is not as tall as Rick.

C. Sally is taller than John but not as tall as Rick.

D. Rick is taller than John and Sally.

2. In spite of the loud music, I soon managed to fall asleep.

A. The loud music soon helped me to fall asleep.

B. I soon fell asleep as a result of the loud music.

C. The loud music made me unable to fall asleep soon.

D. I soon fell asleep even though the music was loud.

3. If you'd forgotten to put out your hand, you wouldn't have passed your driving test.

A. You didn't forget to put out your hand and you passed your driving test.

B. You forgot to put out your hand and you failed your driving test.

C. You forgot to put out your hand but you passed your driving test.

D. You didn't forget to put out your hand but you didn't pass your driving test.

Type 3 This item type consists of a very short reading extract of only a few sentences (or sometimes of only one sentence). The testees are required to answer only one comprehension test item on each reading passage. The actual construction of multiple-choice reading comprehension items based on a reading extract will be treated in greater detail in the next section. Meanwhile, here are two examples of the use of multiple-choice items for testing reading comprehension, the first being at a fairly elementary level and the second at a more advanced level.

1) The president was talking to a young woman in the crowd when Bill suddenly caught sight of a man standing several yards behind her. The man had something in his hand: it was a short stick.

What made Tim notice the man in the crowd?

A. He was very close to Tim.

B. The president was talking to him.

C. He was standing in front of the woman.

D. He was carrying a stick.

2) There were only two ways of moving along the narrow ledge: face outwards or face to the wall. I concluded that even the smallest of bottoms would push a person with his back to the wall far enough out to overbalance him and so, with arms outstretched in the shape of a cross and with chin pointed in the direction I was heading, I inched my way along.

The writer managed to cross the narrow ledge by

A. crawling along on his knees with his arms stretched out in front c him.

B. moving sideways inch by inch with his back to the wall.

C. working his way forward on his stomach with his face almost touching the ledge.

D. walking slowly with his face and stomach close to the wall.

2. Longer Texts

The multiple-choice test offers a useful way of testing reading comprehension. However, not all multiple-choice reading tests are necessarily good tests of reading comprehension. As was clearly indicated earlier, the extent to which a test is successful in measuring what it sets ouT to measure depends largely on the effectiveness of each of the items used. Indeed, certain general aspects of many reading tests may be suspect. For instance, does the usual brief extract for reading comprehension concentrate

too much on developing only those skills required for intensive reading, encouraging frequent regressions and a word-by-word approach to reading?

The sampling of the reading passage is of the utmost importance and must be related to the broader aims of the language teaching situation. Many of the texts in both school and public examinations concentrate too much on a literary kind of English. If certain students are learning English in order to read technical journals, for example, then the sampling of the reading extract should reflect this aim. Ideally, in a test of proficiency the text should contain the type of reading task which will be demanded of the testees in later real-life situations. If the test is a class progress or achievement test, the reading passage should be similar to the type of reading material with which the students have been confronted in their work at school. In other words, if other subjects are being taught in the medium of English (as in many second language situations), the text should frequently (though not always) reflect the type of reading the students are required to do in history or chemistry, etc.

In this section, it is assumed that only intensive reading skills are being tested. Thus, the length of the reading extract recommended might vary from 50 to 100 words at the elementary level, 200 to 300 words at the intermediate level, and 400 to 600 words at the advanced level. These figures are, of course, extremely rough guides and may not be appropriate for many reading situations. Moreover, the extract selected should be capable of providing the. basis for a sufficient number of multiple-choice comprehension items. It is not an easy task to find an extract which will

support a number of multiple choice items - even though the same extract may form a basis for a large number of true/false items or open-ended questions. Generally speaking, passages dealing with a series of events, a collection of facts, or different opinions and attitudes make the best types of texts for testing purposes; those dealing with a single idea or main thane are rarely suitable.

The length of the extract should also be related to its level of difficulty: a particularly difficult or complex passage would probably be considerably shorter than a more straightforward one. On the whole, the difficulty level of the text, however, should coincide with the level of the students' proficiency in English, but we must remember that the reading matter used outside the test situation (e.g. simplified readers) should be selected for enjoyment and should thus be at a slightly lower level than the actual standard of the reading skills acquired. (The difficulty level of a text depends chiefly on the degree of the structural and lexical complexity of the language used.)

When writing test items based on a reading text, the tester should attempt to construct more items than the number actually required. After the construction of the items, it is useful to secure the services of one or two colleagues so that all the items can be moderated. Invariably this process brings to the attention of the item writer certain flaws in some of the items. Although a number of the flaws will be easily rectified, in certain cases it will be necessary to dispense with entire items. In tests of grammar and

vocabulary, new items can always be constructed in place of the discarded items, but this does not follow with reading comprehension items. The text itself has to be rewritten, certain sections added and others deleted in order to obtain the required number of items. Such processes are difficult and time-consuming: thus, it is always an advantage to construct in the first instance more items than are actually required. If the text will not allow for more items, another more suitable text should be chosen to avoid wasting time at a later stage.

How many multiple-choice items should be set on one text? Clearly, the number of items will depend on the length and complexity of the text. However, tests of reading comprehension generally contain fewer items than other skill tests. Furthermore, the testees require much more time: to work through a reading comprehension test since they first have to read the text carefully once or twice before they can begin to answer the items based on it. While as little as ten or fifteen seconds for each item can be allowed in multiple-choice tests of. grammar and vocabulary, at least one or two minutes must be allowed for each item in the average reader test, (if the time required to read the text is taken into account). Consequently, such tests, though long in terms of time, must of necessity be short in terms of items and, therefore, less reliable.

The construction of items depending simply on a matching of words and phrases should be avoided. Items should test more than a superficial understanding of the text and should require the testees to digest and interpret

what they have read. The following examples show how ineffective items can be if testees are simply required to match the words in the items with the words in the text.

At four o'clock on September 30th two men armed with iron bars attacked a soldier in Priory Street.

What happened at four o'clock on September 30th?

A. Two nemesis Deraden with Rinot babblers tacklened a derisoldt.

Imagine that a testee did not understand much of the sentence to the text. In order to appreciate this fully, it is necessary to change the situations slightly, and the text might appear to like this:

At four o'clock on September 30th two nemesis Deraden with Rinot babblers tacklened a derisoldt.

What happened at four o'clock on September 30th?

A. Two nemesis Deraden with Rinot babblers tacklened a derisoldt. Etc

A slightly better item stem would be:

What happened one afternoon at the end of September?

However, to be completely satisfactory, it would be necessary to rewrite both the text and the item, as in the following example:

Tom was surprised when he met Sue at the party. He was under the impression she had gone away from the locality. The last time he saw her was when Jane was teaching her to drive. A few days afterwards she had suddenly become ill.

(first version)

Tom was surprised when

A. Sue went away.

B. he met Sue at the party.

C. Jane was teaching Sue to drive.

D. Sue suddenly became ill.

(second version)

Paul did not expect to see Sue because

A. he knew she was at the party.

B. he thought she had left the district.

C. he had seen Jane teaching her to drive.

D. he had heard she was ill.

There is often a temptation to concentrate too much on facts, figures and dates when constructing test items based on a factual text. Generally speaking, figures and dates are included in a text chiefly for the purpose of illustration or to show the application of a general principle. It is useful in such cases to construct items which require the testees to use the figures in the text to state (or restate) the general principle behind them.

Example: From January to December last year, 291 people were killed and 6,248 were injured in road accidents on the city's roads. 157 of all the fatal accidents involved motorcyclists or their pillion passengers, while 95 _ involved pedestrians and the remaining 39 the drivers and passengers of motor vehicles.

Over half of all the people killed in road accidents last year were

A. motorcyclists and pillion passengers.

B. pedestrians.

C. drivers of buses, cars and lorries.

D. both pedestrians and motorists.

Testees can also be encouraged to use the figures they are given in a text and to work out simple arithmetical sums and problems. Clearly, there is a limit to the tasks which the testees may be required to perform: otherwise the test writer will be testing something other than language skills. The following is an example of an item which tests students' ability to handle simple facts and figures in English: the stem presents a useful task provided that this kind of reading exercise is not overdone.

Latest reports from the northeast provinces state that at least sixteen people lost their lives in Saturday's floods. A further nine people, mostly children, are reported missing, believed dead. Seven small boys, however, had a miraculous escape when they were swept onto the branches of some tall trees.

The total number of people reported dead or missing as a result of Saturday's floods is :

A.7 B.9 C.16 D.25 E.32

The choice of the-correct option in each multiple-choice item must depend on a testees comprehension of the reading text rather than on his general knowledge or intelligence. The following item, for example, can be

answered without any knowledge of the text on which it has been based.

Memorizing is easier when the material to be learnt is
A. in a foreign language.

B. already partly known.

C. unfamiliar and not too easy.

D. of no special interest.

Care must be taken to avoid setting distracters which may be true, even though they may not have been explicitly stated by the writer. In the following test item based on a reading text about the United Nations and the dangers of war, C is the required answer; however, all four options are correct - even though not stated in so many words by the writer.

What would happen if there was a global war?

A. Nations would train men for war.

B. Lots of terrible weapons would be made.

C. The whole human race would be completely destroyed.

D. People would grow very desperate.

The correct option must .be roughly the same length as the distracters. In the following test item the correct option has been modified to such a degree. That it appears as the obvious answer without even necessitating any reference to the text.

The curriculum at the new college is a good one in many ways because it

A. includes many science courses.

B. offers a well-balanced programme in both the humanities and the

sciences.

C. is realistic.

D. consists of useful technical subjects.

All the options must be grammatically correct: there is a tendency especially in reading comprehension to overlook the grammatical appropriateness of some of the distracters used. Option D in the following item can be ruled out immediately because it is ungrammatical.

The writer says that he had studied engineering for

A. a long time.

B. only a very short period.

C. several years.

D. never.

Double negatives are only confusing and such items as the following (based on the extract on page 120) are best avoided:

Paul did not expect to see Sue because

A. he did not know she was at the party.

B. no one knew she had left the district.

C. he hadn't seen Jane teaching her to drive.

D. he didn't realize she was well.

A useful device in multiple-choice tests of reading comprehension is the option ALL OF THESE or NONE OF THESE:

According to the passage, what do some people think there should be outside a modern city?

A. Buses

B. Car parks

C. Office buildings

D. Taxis

E. ALL OF THESE

If an option like E is used, it is advisable to have it as the correct answer in at least one of the items. The testees should not be encouraged to think that it has been included simply to make up the required number of options.

The following text and comprehension items6 illustrate some of the guidelines laid down. in. This section:

Study the following passage and then answer the questions set on it.

“The Captive is a strange but sincere and tender film, as indeed one would expect from a director of the caliber of Marcel Lyme. In addition to his keen sensitivity, Lyme has a strong feeling for historical atmosphere, so apparent in his earlier film Under the Shadow of the Guillotine, in which the events of the French Revolution are depicted with surprising realism and vitality. In The Captive Lyme manages to evoke the atmosphere of an English town in the early part of the nineteenth century, not so much through the more obvious devices of stage-coaches, old inns, and thatched cottages as through minute attention to details of speech, dress, customs, and mannerisms. Similar in theme to Adam Brown, The Captive is distinguished by a sincerity which the former lacks and which helps to transform this film from an

ordinary adventure story into a memorable and a very moving tragedy. Especially unforgettable is the farewell scene at Plymouth, when -Jonathan Robson sees Catherine Winsome on his way to the grim, squalid ship which is waiting to take him to Australia. Robson breaks loose from his captors for a fleeting moment to bid farewell to Catherine. I'll prove my innocence, he cries vehemently as he shakes his fist at Catherine's cousin.

As the ship sets sail, one enters a grotesque nightmare world in which evil seems triumphant. Our identification with Robson becomes so personal that we feel every stroke of the flogging after he has been caught stealing medicine for his sick companion. We share his sympathy for Joe Biggs as the old sailor is hauled under the ship's keel. Indeed, events might well have become unbearable but for the light relief provided by the comical antics of Bobo, the small cabin boy who skips about uncomplainingly doing whatever task he is given. We know, of course, that ultimately evil will be vanquished, and so we are given strength to endure the adversities which confront the hero. The mutiny and the consequent escape of Jonathan Robson, therefore, come as no surprise.

Questions

a. For each of the following statements choose the word or phrase that best completes the statement according to the information contained in the passage. Write the number of the question and the answer you have chosen in your answer book.

i. The Captive was directed by

A. Jonathan Brown.

B. Adam Brown.

C. Marcel Lyme.

D. Catherine Winsome.

ii. In The Captive Marcel Lyme conveys the atmosphere of the nineteenth century chiefly through

A. close attention to small details.

B. the use of conventional scenery.

C. stage-coaches, old inns, and thatched cottages.

D. depicting dramatic events of the time.

iii. The passage implies that Adam Brown was

A. a very moving film.

B. a realistic and vital film.

C. an ordinary adventure film.

D. a sincere film.

iv. Jonathan Robson is angry as a result of

A. having to wait to go to Australia.

B. being wrongly convicted.

C. meeting Catherine.

D. being recaptured.

v. On the voyage to Australia Robson

A. becomes ill.

B. begins to have nightmares.

C. is hauled under the ship's keel.

D. receives a flogging.

vi. Bobo is introduced into the story to help us to bear the grim events by

A. behaving in a strange but interesting way.

B. making us laugh.

C. doing everything without complaining.

D. acting kindly toward the hero.

vii. We can endure the hero's sufferings because we know

A. things cannot get worse.

B. the crew will mutiny.

C. good will win in the end.

D. the hero is very brave.

viii. The writer's attitude to this film is

A. appreciative

B. patronising.

C. scornful.

D. critical.

ix. The word 'his' in line 3 refers to

A. 'The Captive' (line 1)

B. 'one' (line 1)

C. 'a director' (line 2)

D. 'Lyme' (line 3)

x. The words 'the former' in line 12 refer to

A. 'theme' (line 11)

B. 'Adam Brown' (line 11)

C. 'The Captive' (fine 11)

D. 'a sincerity' (line 11)

xi. The word 'his' in the phrase 'We share his sympathy' in line 23 refers to

A. 'Robson' (find 21)

B. 'his sick companion' (line 23)

C. 'Joe Biggs' (fine 24)

D. 'the old sailor' (line 24)

xii. The word 'he' in line 27 refers to

A. 'Robson' (line 21)

B. 'Joe Biggs' (line 24)

C. 'the comical antics' (line 26)

D. 'Bobo' (line 2.6L_

b. Each of the following words and phrases can be used to replace one word in the passage. Find the words and write them in your answer ook. Number your answers.

i. Dragged

ii. Conquered

iii. troubles and misfortunes

iv. very brief

v. finally

I. Curriculum and Syllabus

As it is known that now the government targeted the school should use *KTSP*. *KTSP (Kurikulum Tingkat Satuan Pendidikan)* or School-Based Curriculum is curriculum targeted by the government started from 2006 replacing the *KBK (Kurikulum Berbasis Kompetensi)* targeted by the government in 2004. In *KTSP*, the government allows the teachers or the members of committee of each school to arrange and improve the curriculum or syllabus by their selves under the coordination from the regency. To help the school and the regency, the government gives guidance to approve the curriculum and the syllabus.

Furthermore, in formal education, a curriculum (plural curricula) is the set of courses, and their content, offered at a school or university.[32]

A syllabus is an outline and summary of topics that are covered in a course. A syllabus usually contains specific information about the course, such as information on how, where and when to contact the lecturer and teaching assistants; an outline of what will be covered in the course; a schedule of test dates and the due dates for assignments; the grading policy for the course; specific classroom rules; etc.[33]

[32] Wikipedia, *Curriculum*, http://en.wikipedia.org/wiki/Curriculum, September 15, 2013.
[33] Wikipedia, *Syllabus*, http://en.wikipedia.org/wiki/Syllabus, September 15, 2013.