Reading Test Construction

B. Budiyono
B. Himawan Setyo Wibowo

Abstract: The importance of reading skills accounts for a number of credits to reading courses. It also accounts for a variety of formats of assessment such as weekly assessment and end-of-term tests. For that reason, Barrett's taxonomy proposes itself to be a reference for reading test construction. It suggests operational skills to be tested.

Keywords: recall, inferential, evaluation

Introduction

The importance of reading skills is more explicitly reflected in a great number of credits to reading courses intended to develop effective reading that is motivated by a purpose and requires the ability to judge what information is needed, what degree of comprehension is requested, how long the information will have to be retained and the ability to adapt the speed and strategies (Ferguson, 1973: 23-31). It accounts for various formats such as multiple-choice and matching tests, or subjective such as essay tests, or other formats such as cloze procedure, completion or short answer tests (Heaton, 1989), which can be formulated in reference to Bloom's (1956) cognitive domain, which is a general educational reference, or to Barrett's (1974) content area reading.

Purpose of the Study

This study was intended to develop a test to measure achievement in Reading B at the Department of English at Widya Mandala Catholic University in Surabaya by means of a multiple-choice test.

Review

1. Reading

Reading is usually assumed to be of the following models. In the bottom-up model, reading is believed to be a decoding process of reconstructing the author's intended meaning from the smallest textual units at the bottom (letters and words) to the largest units at the top (phrases, clauses). Reading is deriving meaning from print (Adams 1980:15; Askov 1982: 19). The top-down model has developed from Goodman's psycholinguistic model of reading (Carrell and Eisterhold, 1988:74), where a set of expectations and predictions are made (Eskey 1993:2-3). The interactive model assumes that skills at all levels are interactively available to process and interpret the text, as proposed by Weber (1984) in Grabe (1988:58). Background knowledge and various

types of language knowledge interact with information in the text to contribute to text comprehension (Weir, 1993: 64). The schema theory model emphasizes the importance of background knowledge in comprehension (Rumelhart, 1980:34). Schemata are textual knowledge structures used by a reader to understand a given text (Wolf 1987:309). They are activated by the linguistic cues of the text and the reader's expectations. Once they have been activated, they are used as guiding structures in comprehension. They are classified into linguistic schemata, content schemata, and formal schemata (Carrel and Eisterhold 1988:79).

To assure comprehension, different types of questions are suggested. Barrett (1972), for example, classifies comprehension questions into these types: literal, inference, evaluation, and appreciation. Literal comprehension requires the recognition or recall of ideas, information, and happenings that are explicitly stated in the materials read. Recognition tasks, which frequently take the form of purposes for reading, require the student to locate or identify explicit content of the reading selection. Recall tasks demand the student to produce from memory explicit statements from a selection; such tasks are often in the form of questions teachers pose to students after a reading is completed. Inferential comprehension is demonstrated by the student when he uses a synthesis of the literal content of a selection, his personal knowledge, his intuition and his imagination as a basis for conjectures or hypotheses. Generally, then, inferential comprehension is elicited by purposes for reading, and by teachers' questions which demand thinking and imagination which are stimulated by, but go beyond, the printed page. Evaluation is demonstrated by a student when he makes judgments about the content of a reading selection by comparing it with external criteria, e.g., information provided by the teacher on the subject; authorities on the subject, or by accredited knowledge, or values related to the subject under consideration. In essence, evaluation requires students to make judgments about the content of their reading, judgments that have to do with its accuracy, acceptability, worth, desirability, completeness, suitability, timeliness, quality, truthfulness, or probability of occurrence. Appreciation has to do with student's awareness of the literary techniques, forms, styles, and structures employed by authors to stimulate emotional varying degrees of inference and evaluation, but their primary focus must be on intellectual impact on their readers.

In comparison to Bloom's taxonomy, the first three of Barrett's levels of comprehension are closely related to Bloom's cognitive levels, Barrett's *inferential* to Bloom's *application* and *analysis* and Barrett's evaluation to Bloom's *synthesis* and *evaluation*.

These types of questions appear in a variety of formats, e.g., matching, true-false, multiple-choice, completion, essay, and cloze procedure (Heaton, 1989). True-false items are used to test comprehension of broad principles, application as well as factual details. In this type, trivial, negative and complex statements should be avoided, the equal length of statements of be maintained (Gronlund, 1981:167-169). Matching test items occur in clusters of premises, responses and directions for matching these two. The premises and responses should be homogenous, relatively short, arranged alphabetically, and on the same page (Ebel (1979:173-174; (Gronlund 1981: 175).

Multiple-choice Items are those in which the lead or stem is followed by two to five responses only one of which is usually correct. Multiple-choice items are the most highly regarded and widely used form of objective test items and are adaptable to the measurement of most educational outcomes e.g. knowledge, understanding, judgement, ability to solve problems, to recommend appropriate actions and to make predictions. Almost any understanding or ability that can be tested by means of any other item form can be tested by means of multiple choice items. Gronlund (1981: 189-198) has proposed 13 principles in the construction of multiple choice items.

In addition to the objective types above, there are essay tests. Essay tests are said to provide clues to the nature and quality of test takers' thought processes, critical thinking, originality and idea organization (Ebel, 1979: 96). Gronlund (1981:229-233) has proposed 5 principles for the construction of essay tests.

Another format is referred to as cloze procedure working through the deletion of every n-th word, usually between every 7^{th} and 10^{th} word (Djiwandono, 1996: 80), certain words such as those of proper nouns and dates (Oller, 1979: 346). There are two principal methods of scoring. The exact-word method counts as correct the answer that is exactly the same as that originally found in the text. The contextually-appropriate word method counts as correct the answer that is suitable with the context in terms of meaning, form and grammar (Oller, 1983: 207).

2. Test Development

Stages

Test development is organized into three steps: design, operationalization and administration (Bachman and Palmer, 1996: 85-93). The design stage details the components of the test design to ensure that the performance on the test tasks will correspond as closely as possible to language use and that the test scores will maximally be useful for their intended purposes. This stage includes the description of the purpose of the test, the task

types, the test takers. It also involves the definition of the construct to be measured, a plan for test evaluation, and management of available resources.

The operationalization stage involves the production of test task specification and a blueprint, i.e., how the actual test is to be constructed. It also includes writing the test instructions (the structure of the test, the tasks and how the test takers are expected to respond to the test. This stage also specifies the scoring method (the criteria for evaluating the test takers' responses and the scoring procedures).

The third stage is test administration. It involves preparation of the testing environment, giving the test to the test takers and collecting the test materials. This stage also includes analyzing the test scores (description of the test scores, item analysis, reliability estimation and investigation of the validity of test use).

Gronlund (1985: 123-253) suggests a linear sequence of test development stages: planning, administrating and scoring, and appraising. Planning involves the statement of the purpose of the test, test specification, selection of item type and the preparation of the test items. Administrating and scoring involve the assurance of giving a fair chance for achievement demonstration during the test and consideration of the availability of separate answer sheets and a scoring key. Appraising deals with the evaluation of the test through item analysis. Appraising makes Gronlund's (19850 sequence more than Bachman and Palmer's (2000). Furthermore, operationalization seems to be part of test planning.

Scoring and Grading

Scoring means the process of correcting tests in reference to the key answer and giving points (the raw scores) to the correct responses (Finocchiaro and Sako, 1983: 308; Lien, 1980: 328). The scores in multiple-choice are obtained by giving one point for each correct answer. The scores are then conversed into the 0-to-100 scale. The next step is grading. It involves comparing a test taker's performance with that of his classmates (norm-referenced grading) through the computation of the mean and standard deviation (Djiwandono, 1996: 124 - 125). Grading may also be criterion-referenced, i.e., comparing a test taker's performance to a prescribed standard that is typically concerned with the degree of mastery to be achieved and the percentage of correct answers to be obtained on a test. It requires a clearly defined domain of learning tasks, a clearly specified and justified standard and a criterion-referenced measure of achievement (Gronlund, 1981: 524-527). This grading may be assigned on the basis of a percentage system. For example (Lien, 1980: 330), 93-100 will be assigned A (excellent), 85-92 (B, very good), 78-84 (C, good), 70-77 (D, poor) and below 70 (F, failure).

Item Analysis

Item analysis is a means of measuring to what extent any single task (item) contributes to the information about an individual provided by a test as a whole. Item analysis is generally measured from the point of view of the content of the item and of its performance. For this purpose, an item analysis would result in an item difficulty index (from *too easy* to *too difficult*) and an *item discrimination index* (from *poor* to *very good* (Gronlund, 1981: 258; Hopkins and Stanley, 1990: 269-274).

The contribution of item analysis to the objective type tests is clear. Item analysis provides quantitative data in the selection and improvement of test items. Items of the middle difficulty level and the highest discrimination value should be selected. Those that are too difficult or too easy and have a low level of discrimination should be revised. It is also necessary to arrange test items in order of difficulty, so that the examinees begin with the confidence and reduces the likelihood of their wasting too much time on items beyond their ability to the neglect of easier items they can correctly complete.

Qualities to be Considered
(Finocchiaro and Sako, 1983: 25-31)

The first quality to be considered is validity, the degree to which a test measures what it is designed to measure. The validity is always specific to the purpose of the test. There are different kinds of validity, i.e., content validity, concurrent validity, predictive validity, construct validity and face validity. The second quality is reliability, which refers to (1) the accuracy of a test imeasuring consistently what it is supposed to measure. It also refers to (2) a general quality of stability of scores regardless of what the test measures and (3) the consistency of a measure upon repeated administrations (within a brief interval) to the same group of students. There are different kinds of reliability, i.e., test-retest, alternate form, split-half method, and rater reliability (by the same person on different occasions (intra-rater reliability) or by different people (inter-rater reliability).

There are formulas for reliability coefficients (Ebel, 1979: 279-280). The coefficient of correlation between scores on two reasonably equivalent halves of a test can be computed by the Spearman-Brown formula. The other formulas are Kuder-Richardson 20 and Kuder-Richardson 21.

Kuder-Richardson 20:

$$r = \frac{k}{k-1}\left[1 - \frac{\Sigma pq}{\sigma 2}\right]$$

where
k is the number of items in the test
Σ is the symbol of the sum
p is the proportion of correct responses to a particular item
q is the proportion of incorrect responses to that item (p+q=1)
σ2 represents the variance of the scores on the test

Kuder-Richardson 21:

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

k is the number of items in the test
M is the mean
σ2 represents the variance of the scores on the test

Kuder-Richarson 20 requires information the proportion of correct responses of each item in the test. If the items do not vary widely in difficulty, Kuder-Richardson 21 is suggested. It requires the scoring of 1 for the correct response and 0 for the incorrect one (Djiwandono, 1996: 101).

Another is quality is practicality, i.e., the feasibility or usability of a test in the situation for which it is intended; e.g., is it inexpensive, quickly scored, and easily interpreted

Methods

The methods in this test development project are based on the stages by Gronlund (1985) without avoiding some details suggested by Bachman and Palmer (2000).

A. Planning

Purpose of the Test

The test was intended to measure the test takers' achievement in Reading B, i.e. comprehending main ideas and supporting details of texts taken from magazine, newspapers and journals, as formulated as the objective of reading B in the course outline (p. 138). It is a summative test with the focus of measuring a broad sample of the course objective, to be administered at the end of the course, with the **norm-referenced** type of instrument.

Specification of the Test

The test is specified by its target language use to be measured, i.e. comprehension, which is measured by the test takers' ability in answering comprehension questions as outlined in Barrett's (1974) taxonomy. This specification considers Reading B instructors' suggestion to focus on inferential questions of which

the answers can not be given by the test takers without recall of the explicitly stated information. This suggestion is very helpful because recall questions are likely to be very easy for pre-advanced Reading B takers and such very easy questions have to be eliminated to enhance item discrimination.

Characteristics of the Test Takers

The test takers were the students of the Department of English at Widya Mandala Catholic University in Surabaya taking Reading B, male and female students who had completed Reading A, reading course with intermediate level textbooks such as Dubin and Olshtain's (1981) *Reading by All Means* and Spargo and Glenn's (1980) *Timed Reading – Book Five.* Another prerequisite to Reading B is intermediate *Structure B.*

Selection of the Item Type

Multiple-choice items were developed for efficiency in measuring a large number of question types and a large body of text content. This type avoids the influence of writing skills and of the time for writing the answer (in essay type) and, therefore, maximizes construct-relevance. The other reason is that this type is likely to result with high reliability because of objective scoring. The number of the test takers and the limited time test administration and test evaluation also account for this type.

Preparation of the Test Items

1. Materials Selection

A sample of six texts were taken from magazines and a guidebook to obtain authentic texts (text sources as stated in the course outline) after prior reading the texts already presented in the classroom. They were consulted to Reading B instructors and three were recommended for test development. They were judged to compare to the level of difficulty of the texts already presented. The three texts were "Progress in the Fight against Breast Cancer" from *Voice of America,* April/May, 1988, p. 8, "Computing Random Thoughts" from *Asiaweek* March 2, 1994, p. 41, and "What Americans are Like" from the *Pre-departure Orientation Handbook*, pp. 103-105.

2. Test Items

Every test item was intended to measure one of the types of reading comprehension questions. There were several items for one question type because the number of the items was greater than the number of sampled question types.

Table of Specification

	Question Types	Levels of Importance		N
		Main Ideas	Supporting Details	
1	Recall	11;17;24;26;35	3;4;10;12;15;19;20;29; 33;40;44	16
2	Inferential	6;8;13;18;22;3 4;41;47;50	1;2;5;14;16;21;23;25; 27;28;31;32;36;37;38; 42;43;45;49	28
3	Evaluation	7;9;30;39	46;48	6
N		18	32	50

The cross-sectional table contains the two levels of importance (main ideas and supporting details) that are explicitly stated in the objective of Reading B and three types of questions as suggested in Barrett's (1974) taxonomy. The table also shows the distribution of 50 test items within six cells. The table reads sideways that there are 16 items of recall questions, 28 items of inferential questions and 6 items of evaluation questions. It reads downwards that there are 18 questions for the main ideas and 32 items for questions of supporting details.

3. Constructs to be Measured

1. Reading comprehension: Interaction between the language of the text and the reader's language proficiency and knowledge of the world for the purpose of literal, inferential and evaluative comprehension.
2. Literal comprehension: The ability to answer questions that require recall of ideas, information, and happenings that are explicitly stated in the materials read
3. Inferential comprehension: The ability to answer questions that require a synthesis of the literal content of a selection, personal knowledge, intuition and his imagination as a basis for hypotheses not explicitly stated.
4. Evaluation comprehension: The ability to answer questions that require judgments about the content of the selection.
5. Main Ideas: The ideas assuming the top hierarchy, considered to be of the highest level of importance in a paragraph, either explicit or implicit.
6. Supporting Details: The ideas assuming hierarchy lower than that of the main idea and summarizable into the main idea in a paragraph.

Practicality and Other Qualities:

The test is feasibly developed and administered to Reading B takers at the appointed time and place. Reading B instructors are willing to proctor the test without any test fee. The multiple-choice test is inexpensive, quickly scored, and easily interpreted.

B. Test Administration

The test was administered to two groups of test takers. There were 53 test takers in these groups taking the fifty-item test within sixty minutes. The time was the result of a compromise between fifty minutes as suggested by the test constructor and seventy-five minutes as suggested by reading instructors. They reasoned that the number of the texts, the level of difficulty as compared to the target test takers' competence and the number of test items would be really challenging. Insistence on fifty minutes, i.e., keeping the 1-to-1ratio of time-to-item as suggested by Madsen, 1983: 90), would have been judged to be overrating.

Before the test, the proctors (two reading instructors) explained the purpose of the test, the number of texts, the number of test items, the type of test and the time. After the test was over, the proctors submitted the test papers and the answer sheets to the test constructor and reported test anxiety because of the number of the texts (3), of the test items (50) and of the test pages (eight problem sheets and one answer sheet). They also reported this controversy: that the test takers told them that the texts were not too difficult but they needed more time.

A mild degree of anxiety was tolerated because it is believed to enhance or facilitate performance (Ebel, 1979: 183).

Correction and scoring were computerized by assigning one point to the answer that matched the key and zero to the answer that did not match the key. The scores were interpreted in terms of norm-reference to be graded the way suggested by Djiwandono (1996: 124).

Item analysis was also computerized with the interpretations of item difficulty and discrimination as suggested by Hopkins and Stanley (1990: 269-274). The computer also processed the split-half validity coefficient and the reliability based on K-R 20 and K-R 21.

Results

Test Scores and Grades

The scores are the number of the answers that match the key and computerized by assigning one point to each of those right answers. The grades are assigned based on the score intervals suggested by Djiwandono

(1996: 124) with the computation of the scores, the mean and standard deviation.

Table Scores and Grades

	Grade	Range of Scores	N	%	Normal Curve
1	A	> 31.689	3	5.66%	7%
2	B	25.401 - 31.689	4	7.54%	24%
3	C	12.825 - 25.401	36	67.77%	38%
4	D	6.537 – 12.825	9	16.98%	24%
5	E	< 6.537	1	1.88%	7%
Number of test takers			53	99.83%	100.00%

The table shows a comparison of percentage of the number of test takers with each grade based on Djiwandono's (1996: 124) grading with the percentage of each grade as expected by the normal curve (Gronlund, 1981: 525). The analysis shows no normal distribution.

Discrimination Analysis

Discrimination analysis was computerized by dividing the test takers into two halves, fifty percent for each, referred to as the upper and lower groups. The qualification is as suggested by Hopkins and Stanley (1990: 269-274).

Table Index of Discrimination

	Qualification	Items	N	%
1	Very good	9;29	2	4%
2	Good	3;7;9;13;15;17;43	7	14%
3	Reasonably good	1;4;6;10;18;21;22;27;30;33;34;35;36;37;38;41;46;47;48;50	20	40%
4	Marginal	2;5;14;16;25;26;28;32;40;49	10	20%
5	Poor	11;12;19;20;23;24;31;39;42;44;45	11	22%
Number of test items			50	100%

Ebel (1979: 267) argues that test items of the reasonably good level are *possibly* subject to revision. Hopkins and Stanley (1990: 269-274) argue that test items of the marginal level are subject to revision. The suggestion includes the following points: (1) the reasonably good items should be revised if after a second try-out they prove to be of the same level, (2) the marginal level items should be revised, and (3) the poor level items should be rejected. In so far as there is no second try-out, the decision includes the revision of the marginal level items and the rejection of the poor level items.

Item Difficulty Analysis

The item difficulty index and qualification were computerized based on Hopkins and Stanley's (1990: 269-274) categories.

Table Index of Difficulty

	Qualification	Item	N	%
1	Very difficult	23;27;31;37	4	8%
2	Difficult	5;14;22;33;34;35;38;39;40;44;49	11	22%
3	Medium	1;2;6;7;8;9;10;11;12;13;15;16.17;18;20;21;24;25;26;28;29;30;32;36;41;42;43;45;46;47;48;50	32	64%
4	Easy	3;4;19	3	6%
5	Very easy	-	-	
Number of test items			50	100%

The distribution is less than ideal with the presence of very difficult items (8%) but they may be argued to be worth attempting by top scorers. In general, the distribution is desirable because most of the items belong to the medium level, some to the difficult and easy levels, and none to the very easy level (Ebel, 1979: 273). The very difficult items should be retained to measure a representative sample of reading skills (Gronlund, 1981: 262).

Validity

1. **Content Validity**

   The content validity is attempted at (1) the specification of the course objective of Reading B in chapter 1 as restated in the purpose of the test in chapter 3, and (2) the table of specification of the question types with the number of test items for each type in chapter 3. Three question types and two idea types are cross-sectionally tabulated into six cells of 50 test items (Table 3.1 in 3.1.5.2.).

2. **Construct Validity**

   The construct validity is built on the definition of the constructs to be measured in reference to the course objective. They are *six* constructs: reading comprehension, recall comprehension, inferential comprehension and evaluation comprehension, main idea, and supporting detail, as defined in 3.1.5.3 in chapter 3. The construct validity is also built on development of the test according to the stages of test development and the principles of multiple-choice test construction.

3. **Face Validity**

The face validity of the test is built on the layout. First, the sequence of the test components: test tile (course title, course number, credits, semester number, and time), test instruction and the texts, each of which is followed by the test items. Second, the alternatives are listed down. Third, the test copy is clear. Fourth, the test pages are numbered, except the first page.

Reliability

The computerized reliability coefficient is .771 (K-R 20) or .716 (K-R 21) and interpreted in Djiwandono (1996) as high (The high coefficient ranges from .70 to .89).

Practicality and Other Qualities

The practicality of the test lies with the availability of the separate answer sheet, the availability of the well-lit, air-conditioned rooms and the computerizability of scoring and analyzing the scores. Other qualities include the willingness of two reading B instructors to proctor the test and

Item Revision

Item revision is related to item selection. The items of the very good and good levels of discrimination power are readily selected. Hopkins and Stanley (1990: 274) do not suggest the revision of the items of the reasonably good items. Ebel (1979: 267) speculates on the *possibility* of revision. The decision is the rejection of the items of the poor discrimination power and the revision of the items of the marginal discrimination power.

The revision is logically be based on the factors that lead the items to the low discrimination level: the alternatives especially the distracters and the level of difficulty of the items to be revised. The analysis of those factors is presented in the following table. The table includes the levels of difficulty and discrimination power and the possible causes. The revision is, therefore, focused on improving the possible causes.

Table Analysis for Item Revision

	Item	Discrimination	Difficulty	Possible Causes
1	2	Marginal	Medium	Answer D (70%) very obvious; distracter A very unattractive (2%)
2	5	Marginal	Difficult	Distracter B (60%) much more attractive than answer C (21%)
3	14	Marginal	Difficult	Distracter D (47%) much more attractive than answer A (19%)

4	16	Marginal	Medium	Distracter D (6%) very unattractive
5	25	Marginal	Medium	The answer A (51%) too obvious; distracters B(8%) and C(11%) unattractive
6	26	Marginal	Medium	Distracter A(30%) more attractive than answer B(25%); distracter C (25%) equally attractive as answer B (25%)
7	28	Marginal	Medium	Distracter A (6%) very unattractive; distracter C (34%) more attractive than answer B (28%)
8	32	Marginal	Medium	Answer A (55%) very obvious; distracter C (8%) very unattractive
9	40	Marginal	Difficult	Answer B (40%) very attractive; distracter C (2%) too unattractive
10	49	Marginal	Difficult	Answer B (51%) too obvious; distracter C (8%) unattractive

In general, there are two considerations about the table. First, difficult items are likely to lead to good discrimination (Gronlund, 1981: 263). A strategy based on this assumption will be raising the difficulty level, i.e., from the medium to the difficult level for items 2, 16, 25, 26, 32. This may be conducted by making the stem more specific and the alternatives more similar or homogeneous. More specifically this strategy works as follows:

(1) Item 2: Answer D should be made more similar with the distracters (to be less obvious) and distracter A more plausible to be more attractive.
(2) Item 16: Distracter D should be made more attractive.
(3) Item 25: Answer A should be made more similar with the distracters to be less attractive.
(4) Item 26: Answer B should be made less obvious to be chosen by high achievers and distracter A should be more similar with the others.
(5) Item 28: Distracter A should be made more attractive.

(6) Item 32: Answer A should be made more similar to be less obvious, and distracter C should be made more attractive.

Second, the items of low discrimination power should be examined for the presence of ambiguity and clues (Gronlund, 1981: 263). This strategy applies to items 2, 16, 25, 26, and 32.

Ideally, the items of the difficult level should have led to a good level of discrimination. In part, guessing may account for the low level of discrimination, i.e. a few low achievers guessed the answer because the alternatives are similar and, therefore, difficult. Other possibilities are, however, suggested by the distracter analysis.

(1) Item 5: Distracter B is attractive very much to high and low achievers (tricky). It may sound too plausible. All the alternatives should be made more similar, as plausible as the others.
(2) Item 14: Distracter D is very attractive. The same suggestion applies to this item.
(3) Item 40: Answer B is very attractive. To make it less attractive, distracter C should be made as plausible as the answer.
(4) Item 49: Answer B (obvious, too plausible) should be made less obvious by making distracter C more plausible.

Conclusion

Hard efforts have be made on developing Reading B test, from building construct validity and content validity to providing the alternatives for this multiple-choice test. The construct validity (definition of six constructs to be measured) and content validity (cross-section of six cells of 50 test items) may be assuring, the reliability coefficient (K-R 20: .771; K-R 21: .716) may be relieving, but the discrimination reflects the major weakness of the test.

The marginal discrimination power of ten items requires the revision of ten items, especially the distracters and a few answers, and the poor discrimination of eleven items forces the rejection of those items. A number of distracters have been suggested to be made as plausible as the answers. The rejection of the items may justifiable: it reduces the coverage of the ideas to be questioned but does not reduce the coverage of the question types and idea types. It may also be justifiable because the items to be justified belong to the recall and inferential question types and the idea type of supporting details that are much sampled in the test.

References

Anastasi, A. 1982. *Psychological Testing*. New York: Macmillan Publishing Co., Inc.

Barrett's*taxonomy.http://education.wm.edu/centers*/ttac/documents/packets/inferential.pdf.

Bachman, L. F. and A. S. Palmer. 1996. *Language Testing in Practice*. Oxford: Oxford University Press.

Carrell, P.L. and J.C. Eisterhold. 1988 "Schema Theory and ESL Reading Pedagogy" in Carrell, P.L., J. Devine, and D.E. Eskey (Eds.). 1988. *Interactive Approaches to Second Language Reading.* Cambridge: Cambridge University Press.

Djiwandono, M. S. 1996. *Tes Bahasa Dalam Pengajaran*. Baandung: Penerbit ITB

Dupuis, M.M. and E. N. Askov. 1982. Content Area Reading. New Jersey: Prentice Hall, Inc.

Ebel, R. L. 1972. *Essentials of Educational Measurement*. New Jersey: Prentice-Hall Inc, Englewood Cliffs.

Eskey, D.E. 1986. "Learning to Read versus Reading to Learn: Resolving the Instructional Paradox" in *English Teaching Forum* XXI (July 1983): 2-3.

Ferguson, N. 1973. Some Aspects of the Reading Process. *ELT Journal* VIII: 29-34.

FKIP Universitas Katolik Widya Mandala. 2000. *Pedoman FKIP 2000-2001*.

Grabe, W. 1988. "Reassessing the Term Interactive" in Carrell, P.L., J. Devine, and D.E. Eskey (Eds.). 1988. *Interactive Approaches to Second Language Reading.* Cambridge : Cambridge University Press.

Gronlund, N.E. 1981. *Measurement and Evaluation in Teaching*. New York: Macmillan Publishing Co., Inc.

Heaton, J. B. 1989. *Writing English Language Tests*. London: Longman Group Ltd.

Madsen, H. S. 1983. *Techniques in Testing*. New York: Oxford University Press.

Oller, J. W. 1979. *Language Tests at School*. London: Longman Group Ltd.

Rumelhart, D.E. 1980. "Schemata: The Building Blocks of Cognition" in Spiro, R.J., B.C. Bruce, and W.F. Brewer. (Eds.). 1980. *Theoretical Issues in Reading Comprehension.* New Jersey: Lawrence Erlbaum Associates.

Weir, C. J. 1993. Understanding and Developing Language Tests. New York: Prentice Hall.