The Development of A Diagnostic Reading Test of English for the Students of Medical Faculty, Brawijaya University

Indah Winarni
Brawijaya University

Abstract: This paper describes the development of a diagnostic test of multiple choice reading comprehension as an initial stage in developing teaching materials for medical students learning English. Sample texts were collected from all the departments in the faculty. Selection of relevant texts involved the participation of some subject lecturers. Sixty one items were developed from fifteen texts to be reduced to forty items after pilot testing. Face validity was improved. The main trial was carried out to twenty nine students and item analysis was carried out. The test showed low level of concurrent validity and the internal consistency showed a moderate level of reliability. The low level of concurrent validity was suspected to result from the test being too difficult for the testees as the item analysis had revealed

Key words: diagnostic test, reading comprehension, medical school

English at the medical school of Brawijaya University does not share all the common characteristics of English for the students of non-English departments at the tertiary level of education (henceforth EAP) for undergraduates in general. This is in terms of the attention it has received from the controlling authorities at the medical school, which was reflected in the change that was introduced in 1995 when a language laboratory was installed and the class size was reduced from one hundred and twenty five students to twenty to twenty five students. Pigawahi (1999) claimed that the number of contact hours which was previously equal more or less to

35 x 60' in one semester had increased to 100 x 60'. All these changes were brought about notwithstanding the omission of English from the core curriculum at medical schools in the 1994 curriculum that was applied nationwide.

Although the course offers medical reading and oral presentation, the teaching materials have been geared to a TOEFL preparation in response to the policy of the medical school which set the minimum requirement of TOEFL equivalent score of 450 for the school's graduates. This raises the question of the relevance of the EAP course at the medical school with the needs of the students learning English, which is perceived and formulated by Pigawahi (1999) as (1) reading ability, in order to read textbooks of medical concern written in English; (2) speaking ability, in order to communicate in national and international forum; and (3) writing ability, in order to write scientific papers in English.

Iragiliati (1995) seemed to confirm Pigawahi's perception on the importance of English for medical students. She found out that 80.6% (n:31) of the seventh and eighth semester students were of the opinion that English proficiency was a determinant factor in absorbing knowledge from the lectures. Winarni (1999), however, revealed that most of the graduate respondents (75,3%; n=81) admitted they could finish S-1 program without reading any references written in English, which was supported by 52,7% (n=48) of the subject lecturers. This is contradictory to what Brojonegoro (1998) asserts that the aim of EAP for undergraduate students is to read references written in English. In the context of medical students, Brojonegoro must partly base his statement on the consideration that the dissemination of medical science is mostly written in English (Maher, 1986). This should also be the consideration of the controlling authorities so as not to omit English from the curriculum of the medical school, Brawijaya University.

When asked the reason for reading more references written in Indonesian, only 27.6% (n=71) of the graduate respondents admitted that references written in English were difficult to understand (Winarni, 1999). The rest of them said that references written in Indonesian had already provided adequate information, that translated versions were available, and that references written in English were limited in number. The fact that only slightly over a quarter of the respondents said that references in

English were difficult to read could mean that the English proficiency of the rest was so low that they were discouraged to strive to read English versions of the references. This inadequate proficiency is reflected in the fact that the English proficiency of most of the first year students (77,1%; n=124) is on the range of <400 – 475 on TOEFL equivalent score, which is believed to be insufficient for efficient academic reading.

From the previous discussion, it could be concluded that at Brawijaya University the aim of teaching EAP is to improve students English proficiency of the medical students and that them are not to be discouraged to read references written in English. Apart from all the characteristics of EAP for undergraduates that impose constraints on each level of its operation, (Coleman, 1997, Sadtono, 2001), this article suggests that the EAP course at the medical school of Brawijaya University attract students' attention to reading references written in English by designing course materials derived from topics of medical concern. As English instructors might not be familiar with this undertaking, co-operation with subject lecturers (Dudley-Evans,1998) could be established. Prior to the design of the materials, however, the development of a diagnostic test is suggested. The aim of this paper is to describe the development of a diagnostic reading test for the medical students of Brawijaya University, based on which teaching materials will be developed. The diagnostic test was focused on terminal and enabling objectives of reading comprehension.

DIAGNOSTIC TEST

Brown (1996) emphasizes the role of a CRT diagnostic test in language curriculum once an analysis of needs have been carried out and objectives have been set. The tests will help teachers decide which objectives are the most appropriate, which should be discarded and which should be adopted to a certain degree, for the intended students. Brown further asserts that a diagnostic test will help teachers prepare the undertaking of materials development more accurately before investing all the resources needed in developing the materials to teach the objectives. Bachman (1990: 60) refers to diagnostic test as a test that is "designed and developed specifically to provide detailed information about the specific content domains that are covered in a given program or that are part of a general theory of language proficiency"

READING COMPREHENSION

From a number of comprehensive reviews of reading in a foreign language, Weir (1993) defines reading comprehension as "a selective process taking place between the readers and the text, in which background knowledge and various types of language knowledge interact with information in the text to contribute to a text comprehension". Alderson, et al. (1995: 83) mention reading comprehension as one of the examples of "construct" defined as referring to a "theoretical conceptualization about an aspect of human behavior that cannot be measured or observed directly". It has also been argued that reading is a unitary process that cannot be divided into sub-skills. While Munby's (1978) extensive list of micro-skills has given a considerable influence in course design and needs analysis, it has been strongly critisized for its lack of an empirical base and its impracticality (Alderson, 1998, Lumley, 1993). However, reading sub-skills have been widely referred to as a working construct of reading ability in designing a course as well as in test construction. Weir (1993) is of the opinion that it should be possible to focus on reading sub-skills for testing purposes and Lumley (1993) concludes that involving mapping skills from test content could empirically be justified.

TESTING READING COMPREHENSION

Based on the previous discussion, there are basically two views of testing reading comprehension. The first view takes the understanding that reading is a unitary process and thus, testing of reading comprehension should not be based on the view that reading can be divided into sub-skills. The view that reading can be divided into sub-skills leads to the measurement of reading comprehension through the measurement of the reading sub-skills. Literature on testing reading comprehension provides a lot of information on the ways to test this skill. Despite the complicated process it takes to construct objective test of multiple choice type, this type of test has been widely used as a test for reading due to its practicality in scoring. The techniques of constructing the test have also been widely discussed (Tuckman, 1975; Oller, 1979; Heaton, 1989 and Brown, 1995) .

THE TEST DEVELOPMENT

The stages of test development were adapted from Hughes (1990) and Alderson et al. (1995) resulting in several stages comprising (1) planning and constructing the test (2) pilot testing (3) trying out the test (4) describing the test result (5) analyzing the test result. The following will describe each of the stages.

PLANNING AND CONSTRUCTING THE TEST

This section will describe the learners, the sample text, the reading skills, test format and item writing.

The Learners

The subject of this study were the first year medical students of Brawijaya University owing to the fact that the teaching and learning of English was carried out in the first and second semester. There were 176 students taking the English course. The TOEFL equivalent test on their pre-test shows that 8.6% scored >500; 12.5% (476-500); 24.9% (450-475); 46.6% (400-449) and 7.4% (<400). Three of the students scored >600. The try-out was conducted in a class of 29 students.

The Sample Texts

The sample texts in this project were the references written in English assigned by the subject lecturers for the students to read. They were collected from all the departments in the Medical School through a letter written by the Dean for each department to submit any references written in English assigned to the students to read for the purpose of writing English language teaching materials. Out of 25 departments 21 submitted the requested sample tasks. As different departments suggest, the topics of the sample texts varied from those of basic medical sciences of pre-clinical concerns to those of clinical concern.

The types of tasks were also sought from and discussed with some subject lecturers. This also resulted in the selection of materials to find those that were pedagogically suitable and that were suitable for the schemata of the first year students. Some articles on health were downloaded from the Internet to add to the collected sample texts. Despite the out-dated year of publication, *Simplified Nursing*

(Dakin at al.: 1956) was also recommended as the content was considered relevant to what medical students have to study.

The Reading Skills

For its elaborate description and clarity, the list suggested by Alderson (1988: 91) was considered most appropriate for the purpose of developing the present test. Since it was meant for language skills in general, some modification should necessarily be made. The list of the modified reading skills comprises:

a. Understanding information in the text not explicitly stated through making inferences
b. Understanding explicitly stated information: i.e., where the question does not paraphrase the text, or where the same key word/words in both question and text leads the reader to the answer.
c. Extracting salient points to summarize
d. Recognizing indicators in discourse for: introducing an idea, emphasizing a point, explanation or clarification of points already made, etc.
e. Deducing the meaning and use of unfamiliar lexical items through contextual clues
f. Understanding conceptual meaning of quantity and amount
g. Understanding relation within the sentence especially long pre-modification and post-modification by prepositional phrase
h. Interpreting text by going outside it using exophoric reference

This test was meant to concentrate especially on the first and second skills on the list although, to a certain extent, it could not avoid including other skills. Table 1 shows the distribution of the reading skills in the test.

Table 1. Distribution of Reading Subskills

Reading Subskills	Number of Items
A	1,2,5,9,11,12,20,21,24,25,26,30,33,39
B	3,4,6,8,10,13,14,15,16,17,18,19,22,29,31,32,34,36,37
C	23

Reading Subskills	Number of Items
D	28
E	7,38,40
F	35
G	-
H	-

Test Format and Item Writing

To avoid bias towards one particular test format or to one particular type of learner, Alderson. C. J.et al. (1995) recommend that more than one test format be used to test any ability. For the purpose of the present test development, objective test of multiple choice format was used with the consideration that this should only be an initial part of developing a good diagnostic test.

Although Brown (1995: 55) and Heaton (1989:28) suggest four options for an item, the present test developed five options with the consideration that if five good options could be developed, the test would minimize the chances of guessing (Heaton 1989: 26). The second consideration was that it was thought to be more beneficial to design five options in the beginning than four options because in the revision of the test, when necessary , it was considered easier to drop an option than to revise or create a new one. The steps suggested by Heaton and Brown in constructing multiple choice items were found worthwhile and, as far as possible, they had been referred to in the process of constructing the test.

Alderson et al. (1995), suggest that item writers have to begin their writing task with the test's specification. The next step is to find appropriate texts – texts that have the potential of matching the test specification and would likely provide the items expected. Finding the appropriate texts in this study was carried out by inviting two subject lecturers to participate in the selection especially determining the topic that was within the schemata of the first year students. It was found that as Alderson et.al claim, 'searching for texts that have promise' (1995: 43) took some time. Out of thirty texts provided by all the departments, five were selected to produce thirty five items. Seven texts were taken from Simplified Nursing (Dakin et al. 1956); and two were downloaded from the internet. Table 2 shows

how many items could be generated from one text. It will be learned later that after the pilot testing, this number was reduced to forty.

Table 2. Number of Item Generated from the Texts

No	Topic	No of Item
I.	MENTAL ILNESS	5
II	ANGINA PECTORIS	4
III	INFECTION DISEASE	4
IV	DIABETES MELLITUS	4
V	BREAST CANCER	3
VI	ANTIBIOTICS	2
VII	NERVOUS SYSTEM	3
VIII	HEART INFECTION	3
IX	SPROUTS	2
X	CAUSATION	5
XI	PERCUSSION	5
XII	STETHOSCOPE	4
XIII	PARASITE	4
XIV	ENZYME	5
XV	TRACHOMA	5
XVI	OEDEMA	7
	TOTAL	61

PILOT TESTING

It is suggested that pilot testing, a less formal try-out on a group of colleagues be carried out to iron out the main problems before the major trials (Alderson et.al, 1995; Djiwandono, 1996). In the development of the present test, pilot testing was conducted to a group of colleagues – one English instructor and three subject lecturers and two senior students of 1998. In its preliminary form, the test consisted of fifteen passages with 61 items. Seven of the texts were taken from the references written in English provided by Departments of Opthalmology, Public Health, Parasite, Physics and Pathology; five were taken from Simplified Nursing (Dakin et al, 1956) and two of texts were downloaded from. The testees were given two hours but on the average, they finished the test in one and a half.

Table 3. The Scores of Seven Subjects in Pilot Testing

Subject no.	Raw Score	Percentage score
01	41	67
02	39	63
03	41	67
04	39	63
05	40	65
06	41	67
07	34	50

The general impression of the testees indicated a low face validity due to too many passages of various topics. The English instructor refused to finish the remaining five last items for being "too fatigue" having to read and concentrate on too many different texts. It was suggested that the test eliminate the passages with less than five items. The same impression was given by the other two subject lecturers. The other subject lecturer said that the test was too interesting to feel tired in doing it. Based on the general impression of the four subjects, seven passages were omitted from the test resulting in eight passages with forty items. Discussion with the English instructor and two of the subject lecturers resulted in some revision of certain items and additional items in the revised version of the test. This revision was due to more than one possible answer in an item and further revision was due to the omission of unnecessary extra information in one of the texts The improvement of the face validity of the test was done by putting the instruction and the example on one first page. In the preliminary version of the test, the instruction and the example were not separated from Text One and its first two items.

DETERMINING THE MAXIMUM CRITERION FOR THE SUBJECTS' SCORES

In order to determine the basis of the maximum criterion for the scores obtained on the test by the intended subjects, a tryout of the test was done to a group of three students from different classes who scored 577, 560 and 565 respectively, on TOEFL equivalent for their pre-test (at the beginning of the semester). An attempt to do the try-out to three high-

est achievers on TOEFL equivalent i.e. TOEFL score of > 600 failed since it was final exam time. This was also the reason for the small number of the testees doing the try-out to determine the maximum criterion. Care had been taken that in the intended class to which the main trial was conducted, the highest TOEFL equivalent achieved was 523. The scores on the students' test can be seen in Table 4

Table 4 . The Scores of Three Subjects to be Used as the Maximum Criterion

Subjects No.	Raw Score	% Score
01	30	75
02	31	77
03	29	72
Average	30	75

THE TRY-OUT

The test was tried out on December 8, 2000 on a reading class of twenty-nine students. The decision on the class simply due to convenience reasons. The classes were regarded as homogeneous since there was no attempts from the Faculty to categorize the classes according to the students' achievement. As discussed previously, the test contained forty items of multiple choice type with five options. The time allowed was ninety minutes. The texts were selected from.

1. Dakin et al (1956) – eighteen items
2. Mayo Organ Transplant (2000) – five items
3. Cameron et al. (1978) – 10 items
4. Gray et al (1979) – six items

THE TRY-OUT RESULT

The highest score gained by the student was seventy three and the lowest score was thirty three. The criteria of those needing instruction were determined based on the highest score of the representative of the intended group. This criterion was sought from three students who were then asked to sit for the test. The number of correct answers respectively

were: thirty, thirty-one and twenty-nine. The mean of the score, which was thirty, was to be the ideal score. Applying this criterion to the number of correct answers gained by the students, those who scored equal or more than thirty should get the perfect percentage score of on hundred (Table 5). Based on the result of the test, as intended, decision could be made as to which students should benefit from further instruction and to what extent.

Table 5. Converted Score Against Maximum Criterion

No.	Subjects	Raw scores	Ranks	Converted scores
1	21	29	1	96
2	23	25	3	83
3	24	25	3	83
4	26	25	3	83
5	22	24	5	80
6	27	23	6	77
7	19	22	7.5	73
8	20	22	7.5	73
9	25	20	10	66
10	28	20	10	66
11	20	20	10	66
12	13	19	12	63
13	03	18	13.5	60
14	11	18	13.5	60
15	04	1	15.5	56
16	05	17	15.5	56
17	02	16	17.5	53
18	08	16	17.5	53
19	06	15	21.5	50
20	10	15	21.5	50
21	14	15	21.5	50
22	16	15	21.5	50
23	17	15	21.5	50
24	18	15	21.5	50
25	07	14	25.5	46
26	09	14	25.5	46
27	01	13	27	43
28	12	13	27	43
29	15	13	27	43

TEST RESULT ANALYSIS

The following will discus the test reliability and validity and the item analysis of the test

Test Validity

For the purpose of the paper, the discussion on the validity of the test will be focussed on the external validity, i.e. concurrent validity. Concurrent validation involves the comparison of the test scores with some other measure for the same candidates taken at roughly the same time as the test.

The concurrent validity of the test was obtained by correlating the scores of the students with the scores obtained from the TOEFL Equivalent test for their mid semester test conducted a month earlier. Pearson correlation using SPSS/PC program used to examine the concurrent validity of the present test results in coefficient correlation of .350 which is significant at $p<0.62$ level. Referred to the correlation coefficient range, i.e. +1.00 to – 1.00, the obtained correlation coefficient of the test can be interpreted as having a low level of validity.

Test Reliability

Internal consistency of split-half reliability produces a correlation coefficient of 0.63 at the significant level of (p.00). Adjusted index reliability from Spearman - Brown Prophecy formula gives a coefficient reliability of 0.76. Applying the formula of K-R 21, the coefficient correlation obtained is .49. "The degree of reliability we demand in our educational measures depends largely on the decisions to be made" (Gronlund and Linn RL, 1990). In the development of the present test, the result was expected to identify the need for instruction. A coefficient correlation of .49 should fall on the moderate level of category and this gave the chance of the improvement of items quality.

Item Analysis

Item analysis was conducted to the test mainly to analyze the quality of items to avoid ridiculous distractors of the multiple choice type test.

The item facility estimates showed that on the basis of Brown's (1996) estimation, 55 % of the items in the test were to be retained, 30 % were to be discarded for being too difficult, and 15 % were to be discarded for being too easy. On the basis of Oller's (1979) estimation, however, only 13 % of the item were to be discarded (5 % for being too difficult and 3 % for being too easy).

The result of item discrimination showed that 12 items (30 %) were good items, 12 items (30 %) must be improved while the other 16 items (40 %) must be rejected.

The distractor analysis of the present test development is underway.

CONCLUSION

The present test development was a part of needs analysis of medical students of Brawijaya University learning English. The purpose of the diagnostic test was to inform the students as well as the English instructors as to how well the students were in the reading skills being measured by the test.

A good test must be valid and reliable. Test validity can be measured in various ways. As far as this study was concerned, the efforts of validating the test had been carried out through working on internal as well as external validation of the test. Ensuring the content validity of the test started from the stage of planning by consulting the subject lecturers as to which testing materials were suitable for the students. Input on face validity had also been sought by asking colleagues, an instructor of English as well as three subject lecturers to do the test and discuss what they thought about it afterwards. Some revision of the test was done based on the inputs. Even so, it was learned empirically that the estimate concurrent validity of the test was very low. Apparently, them would justify the face and response validity of the test in the pilot testing.

Judged from the IF on the basis of Brown's estimation, the test could be regarded as probably too difficult for the first year medical students learning English. A look at the Item Facility of the test informed us of the level of difficulty of the test. This could confirm that the low level of concurrent validity obtained was due to the test being too difficult. It is interesting to find out however, how the validity of the test might be, given a group of testes, with significantly better English proficiency.

The low estimate of the test validity reflects on how the test was constructed. Judged from the difficulty of the test, it could be interpreted that the proficiency level of the students was not taken into proper consideration. Bias towards the specific texts that the test takers were assumed to be familiar with, needs further consideration.

The internal consistency of the test showed moderate level of reliability. Item analysis had been applied to the test. Apart from its questionable meaningfulness to the CRT referenced test, it was considered worthwhile to do for especially multiple choice items for it was useful in analyzing defective or ridiculous distractors.

It was learned that from the thirty texts submitted, only two were selected for the development of the test. It can be concluded that the references written in English assigned by the subject lecturers for the students to read were not necessarily appropriate for English teaching and learning materials.

REFFERENCES

Alderson, C. 1988. Testing and Its Administration in ESP. In D. Chamberlain and R.J. Baumgardner (Eds). *ESP in Classroom: Practice and Evaluation*. ELT Document 128.

Alderson, J.C. Clapham, C. and Wall, D. 1995. *Language test construction and evaluation* Cambridge University Press

Bachman, L.F. 1991. *Fundamental Consideration in Language Testing*. New York: Oxford University Press.

Brojonegoro, S.S. 1998. Higher Education Development 1995 – 2006: paper presented in the the Second National Conference on NUESP, 4 – 7 May, Ambon

Brown, JD. 1996. Testing in Language Programs. Upper Sadle River, N.J. Prentice Hall Regents.

Cameron, J.R. & Skotronik, J.B. 1978. *Medical Physics*. New York: Interscience Publication – John Willey and Sons, Inc.

Coleman, H. 1995. Undergate ELT: Where Have We Been and Where Are We Going? In Coleman H., Soedradjat, T. M. and Westaway, G (Eds) *Teaching English to Undergraduates in the Indonesian Context: Issues and Development.* (pp. 26-42) Bandung: ITB Press and School of Education University of Leeds.

Dakin, F., Thompson, E.M. and LeBaron, M. 1956. *Simplified Nursing*. Philadelphia, J.B. Lippincott Co.

Departemen pendidikan dan Kebudayaan Direktorat jenderal Pendidikan Tinggi 1994. Himpunan Keputusan Menteri Pendidikan dan Kebudayaan Republik Indonesia tentang Kurikulum Nasional Program Sarjana. (Compilation of Letters of Decree of Minister of Education and Culture on the National Curriculum of Sarjana Program.

Djiwandono. M.Soenardi. 1996. Tes Bahasa dalam Pengajaran. Bandung, Penerbit ITB.

Gray, C.H. & Howorth, P.J.N. 1980. *Chemical Pathology*. London: The English Language Book Society and Edward Arnold Ltd.

Grellet, F. 1990. *Developing Reading Skills*. New York, Cambridge University Press.

Gronlund, N.E. & Linn, R.L. 1990. *Measurement and Evaluation in Teaching*. New York, Macmilan publishing Co.

Heaton, J.B. 1988. *Writing English Language Test*. New York: Longman, Inc.

Hughes, A. 1990. *Testing for Language Teachers*. Melbourne: Cambridge University Press.

Hutchinson,T and Waters, A. 1987: *English for Specific Purposes, A Learning Centred Approach*: Cambridge University press.

Iragiliati, E. and Andreani, S. 1995. *Pengembangan Model Buku Teks, Lembar Kerja Mahasiswa (LKMK). Tes Bahasa Inggeris Kedokteran Fakultas Kedokteran di Indonesia Timur*. An Interim Report. Institute of Teacher Training Malang.

Lumley, T. 1993. *The Notion of Subskills in Reading Comprehension Test*: An EAP Example. *Language Testing Journal* 17 (1): 211 – 235.

Munby, J. 1978. *Communicative syllabus design*. Cambridge : Cambridge University Press

Maher, J. 1986. English of Medical Purposes. *Language Teaching*, 19: 112-45

MayoOrgan Transplant. 2000. Oral Contraceptives and Breast Cancer. http: www//.mayoclinic.com(accessed October 18, 2000)

Munby, J. 1978. Communicative Syllabus Design. Cambridge: Cambridge University Press.

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

Pigawahi, T. 1999. Pengajaran Bahasa Inngris di Fakultas Kedokteran dan Upaya Peningkatan Kemampuan (skill) Mahasiswa untuk Berkomunikasi dalam Lingkungan Akademis. (The Teching of English in the Faculty of Medicine and the Ettempt to Improve the Ability (skill) of Student in Academic Invironment). *A Paper presented in a seminar on An Attempt to Improve Academic Communication Skill in Interrnational Level for the Students of Medical Faculty, Brawijaya University*. Malang: Medical Faculty, Brawijaya University.

Sadtono, E. 2001. Fighting a Losing Battle: English for Academic Purposes in Indonesia. NUCB Journal of Language Cultural and Communication Vol. 3 (pp.45-58)

Tuckman, B.W. 1975. *Measuring Educational Outcomes: Fundamentals of Testing*. New York: Harcourt Brace Jovanovich, Inc.

Underhill, N. 1987. *Testing Spoken Language: A Handbook of Oral Testing Techniques*. London: Cambridge University Press.

Weir, C.J. 1993. *Understanding and Developing Language Test*. Hertfordshire: Prentice Hall International Ltd.

Winarni, I. 1999. Model Alternatif Pengajaran Aplikasi Bahasa Inggris di Perguruan Tinggi dalam Upaya Meningkatkan Mutu Lulusan (Paradigma Baru Berdasarkan Analisis Kebutuhan) (An Alternative Model for the English Application Course to Improve the Quality of Graduates (A New Paradigm Based on Needs Analysis). An Interim Research Report. Malang: Brawijaya University.