

CHAPTER IV

RESEARCH RESULTS

This chapter shows the results of the study. Before giving the evaluation of the test, a description of the achievement test will be provided. After that, the process of test construction in the BARS program will be presented. This will be followed by the analysis of test content to evaluate the content. Finally, the analysis of the observation data and checklists for estimating reliability of scoring and content will be given.

Description of the test

The first year achievement test of BARS program is evaluated in this study. This is a test that the first year BARS students have to take at the end of the semester. The main objective of the test is to assess how much these students have learnt from what they have been taught.

The test includes discrete items which are intended to test sub skills of reading, writing, grammar and vocabulary. However, the test does not assess listening skills or speaking skills. Three first year English teachers who had at least three years experience of teaching developed the test. These teachers are referred to as Test Developer A (TD-A), Test Developer B (TD-B), and Test Developer C (TD-C) in this chapter.

The test contents are chosen from *New Headway* unit One to Five and the vocabulary supplementary book that the students have been taught from. The test has three main sections: Reading, Grammar and Vocabulary, and Writing. The reading section has three

test formats: multiple-choice, true/false and short answer questions. The second section, grammar and vocabulary has eight test formats, most of them are gap-filling items. The writing section has only one essay writing task. The total time allowed for the test is three hours. The test content is summarized in table 2.

Table 2: Summary of test content

Section	Skills and structures	Test techniques and No. of items	Type of scoring	Points for each section
Section A	Reading (1 passage)	Multiple choice (5 items)	Objective	20
		True/false (5 items)	Objective	
		Short Answer questions (5 items)	Subjective	
Section B	Grammar and Vocabulary	Gap-filling (35 items) Multiple choice (5 items) Matching (5 items)	Objective	45
Section C	Writing	Essay (1 item)	Subjective	15

The students had to take the exam in the evening, between 5 pm and 8 pm. First year English teachers and some teaching assistants administered the test. Then, after the exam, the exam papers were scored by the first year English teachers and teaching assistants. The total score for the test is 80 points. The test contributes 80% of the assessment in the course. The other 20 % comes from the results of in-class tutorials during the semester.

The test results decide whether students are able to attend the program in the next year or not. If the student fails the test, the student needs to re-take the examination.

The content of the test will be discussed in more detail when considering content validity and reliability.

Test Construction Stages in the BARS program

The analysis of the interview data is presented by following the seven stages of test construction discussed in chapter II. The seven stages of test construction process are (1) stating the objectives, (2) writing specifications for the test, (3) writing and moderating of test items, (4) informal trialling of the test items, (5) analysis of the results of the trial, (6) validation, and (7) training of examiners and administrators.

Stating the objectives

There are three main objectives reported for this test. These objectives were set by the test developers individually, based on the parts that they have to develop. The objectives of the test mentioned by the test developers were:

- to assess how much the students have learnt of what they have been taught after the whole semester (TD-A, TD-B, TD-C)
- to assess how much students have learnt the language skills: reading and writing sub skills, the grammar structures and vocabulary from *New Headway* (TD-B, TD-C)

- to assess whether the students are able to use and apply what they have learnt in real life communication. (TD-A, TD-C)

The main objective of all test developers was to assess how much the students have learnt after the whole semester. This was a general objective of the test. The second objective, set by TD-B and TD-C, was a specific objective of the test which was based on the course book they used for the first year students. The third objective was set by TD-A and TD-C and it is more likely to be related to language use beyond the course. The researcher gathered from the interview that the above objectives were not set before constructing the test. These objectives mentioned by the participants were given after consideration, and in response to the interview questions. No specific objectives of the test had been stated with the agreement of all the three test developers before constructing the test.

Writing specifications for the test

No test specifications were written formally in constructing this test. Before constructing the test, all test developers just met for a while and discussed about what they are going to include in the test. All decided that the test should cover what they had taught in that semester. Afterwards, TD-A worked separately and TD-B and TD-C worked together on the test construction. They all had the idea of what they wanted to test in their mind but this was not recorded specifically. TD-A developed the writing test and chose to use an essay type question to assess writing ability.

TD-B and TD-C designed reading, grammar and vocabulary sections and chose the test content from both *New Headway* and the supplementary vocabulary book. For grammar and vocabulary section, they listed what is in the course book and chose what should be included in the test. For the reading section, they decided to use an unseen passage for the students. They had in mind what kind of test techniques will be used, how many items will be included in the tests and how they are going to score each items. However, they did not make any decision about the timing, criterial level of performance and the specific marking scheme for reading and writing.

Even though they chose the content of the test and decided what kind of techniques will be used for each section, they did not have detailed and fixed test specifications which clearly shows what will potentially be in the test, what kind of test structure will be used, timing, medium, techniques to be used, criterial level of performance and scoring procedure for the test. As they did not have specific test specifications to guide them, the test becomes very flexible, and can change at any time. As the test content is undefined, the test developers can add or omit what they want. Therefore, the test is not guaranteed to test what it is intended to test if it does not have a specification.

Moreover, while marking the exam papers, there are some weaknesses in scoring the papers, as no scoring procedure was described.

For the test developers, test specifications do not seem to be important and some did not understand the term “test specifications”. This needed to be explained to them in detail in order to help them understand. From their point of view, specifications were perhaps unnecessary because they believed that the course book content was sufficient.

Writing and moderating of test items

The test developers wrote the test items by using what they had planned in their mind rather than what was explicit on paper. As discussed above, TD-B and TD-C worked together and TD-A worked alone. TD-B wrote the reading section and most of the grammar sections whereas TD-C wrote test items for vocabulary and one part of the grammar section.

For the reading section, the reading passage was chosen by TD-C from one *Pre intermediate* exercise book. True/false and multiple-choice items were adapted from the exercise book from which the passage was chosen and the short answer questions were written by TD-B. This test developer wrote five short answer questions of which three required the opinion of the students. There were altogether fifteen items in this section which meant each format has five items. For the multiple choice and true/false items, there was intended to be only one possible response and no other possible responses were allowed. Short answer questions were scored subjectively.

One problem for TD-B who designed the reading test is that she did not teach reading to the first year students. As she said,

“I was asked to design the reading test even though I am not teaching reading skills in this semester. I think it is better if the teacher who teaches reading could design the reading test.”

For grammar and vocabulary section, TD-B and TD-C worked together. These test developers chose to use some multiple choices, matching and many gap-fill items. The bias for gap fill items was obvious as among eight formats, five formats were gap filling. Some of the items were directly adapted from the textbook and some were re-written as new sentences. However, all of them were either directly or indirectly related to the course book. For example, one item for vocabulary section which assessing word formation was

“Fish soup is a _____ (special) of this area.”

This same sentence can be found in page 28, Unit 3 of *New Headway*. The test developer just took it directly from the textbook.

Some items were rewritten as new sentences. For example, one item for vocabulary section which assessing word formation was

“In Yangon, there are two _____ (industry) zones, South Dagon and Hlaing Thar Yar.”

The answer for this item is “industrial”. In *New Headway* Unit 3, page 28, there is the sentence:

“This is the _____ part of my town. There are lots of factories and businesses.”

The answer of this item is also “industrial”. The test developer tried to write a new sentence to fit with the context.

There were altogether 55 items- 25 items for grammar and 20 items for the vocabulary section. This required only objective scoring and one point was given for each correct

item. Here again, the two test developers had some problem in selecting test content and writing test items. As there are six teachers who are teaching *New Headway* in first year, it was a little bit difficult for them to decide what to put in the test as some teachers had not finished all the units and some teachers just taught from the supplementary book and skipped *New Headway*. Therefore, the lessons that the students had been taught and had learnt can be different from one class to another. TD-B suggested that

“We just know about the level of the students that we have to teach and how far we have taught them. We do not know about the other classes. Now, one teacher taught a lot from vocab supplementary book and she wanted to put it into the test as she has taught it. But the other classes have not received or been taught those lessons. Therefore, it would be unfair for the students from the other classes. We try to meet in break time and discussed what the other teachers have taught. Anyhow, when developing the test, it will be better if all six teachers can meet and develop the test together.”

This suggestion clearly showed the reason why it is better to have explicit test specifications. If the test has explicit test specifications, the test will be a fair reflection of textbook and course objectives and fair for all the students.

For the writing section, TD-A decided to use an essay type writing which she thought might be suitable for assessing the students' writing ability. She chose the topic “The advantages and disadvantages of living in a foreign country” to cover what the students had learnt from unit One to Five. The length of the expected response is 150

words. According to TD-A, the writing test requires subjective scoring and it accounts for 15 of the 80 points in the test. The writing will be judged according to the relevancy, organization, expression of ideas and correct grammar and spelling. TD-A had considered how to score the writings but no specific marking scheme was developed.

After writing the test items, test items were moderated by only one test developer. TD-C took the test by herself to see whether the test items are difficult or not, and whether the language of questions is understandable or not. Then, she also requested another teacher (who is not from BARS program) to check the items for clarity and accuracy. TD-C mentioned about the trialling of the test as

“The trialling of the test items is needed and it should be done. After I wrote the test items, I tried to answer the items by myself first to check whether the students will understand the questions and find the answer. This is to avoid the confusion of the items for the student. I also showed them to one English teacher to check the items, whether the questions are okay or not.”

However, no changes were made after the moderation of the items. Apart from TD-C, the other test developers were not interested in moderation of the test items nor thought it necessary. Moderation of items seemed unimportant for them. One test developer said that

“We met before we developed the test but we did not review it or meet again after we produced the test. We just discussed the test before we developed it. For me, I think if we

can discuss before we write the test, we do not need to review the test items. It is not necessary.”

We can see that moderation of the test items seems unnecessary to the test developers. Items were written individually and no changes were made after writing the test items. This clearly showed that the test developers need more awareness about the importance of moderation of test items.

Informal trialling of the test items

According to the literature in language testing field, the items should be presented in the form of a test to a native speaker and then revisions made if necessary. Afterwards, the items should be put together into a test and administered under test conditions to a group similar to those for whom the test is intended to test. However, in this study, the interviews revealed that no such informal trialling took place.

Analysis of the results of the trial

As no informal trailing had been administered, it is clear that there are no results which should be analyzed for reliability of the items and the test.

Validation

There is no institutional validation of the test before administering it. The discussion of content validation of the test carried out as part of this study will follow in the next section.

Training of examiners and administrators

There was no training for the examiners and the administrators before administering the test. The examiners were briefed by the test developers on how to score the test items before marking the exam papers and were provided with key words. There is no handbook which contains the rationale of the test, description of the test, sample items, or explanation of how test scores are to be interpreted.

The test developers gave some comments on the construction of the test. These suggestions are provided below.

Comments on constructing the test

As only three teachers had to design the test, the test developers had to take responsibility in producing an appropriate test for all the students regardless of how far the student had been taught.

TD-C said that

“If all six teachers can produce the test together, there will be less responsibility for us to take, and as these teachers will know the level and condition of their students, we will be able to produce an appropriate test for them.”

This test developer was talking about modifying the test to suit the students rather than basing the test on objectives or specifications. The test developer cared more about producing a test that is appropriate for the students.

Even though the teachers teach all four skills, grammar, and vocabulary, only reading, writing, grammar and vocabulary were tested in the exam. All test developers wanted to assess listening skills but there are some problems such as electricity shortages, limited resources and time constraints. TD-B mentioned that

“Even though there are four skills in Headway, this test only contains two skills for reasons like electricity, time, materials, cassettes, and the equipment and limited resource persons.”

The other factor is that students have been told by the teachers what to study for the test. The teachers gave hints about what will be in the test and some teachers even chose the specific grammar structures for the students to study for the test. Therefore, the students just studied the chosen structures and sat the exam. This is far from the objective of the

test to cover the syllabus and to see how much they have learnt from what they have been taught. TD-B suggested that

“Regarding this test, if the teachers want to know how much the students have learnt from the lessons, they should not give the hints about what will be in the test. Let them study all the units that have been taught. But now, the students just study what the teachers have highlighted”.

As the test is not clearly testing on objectives and there are no explicit specifications, the test content becomes predictable, leading to this negative backwash. As the teachers are just teaching for the exam, it can lead the students to just focus on the exam and study only to pass the exam.

Content validity of the achievement test

Content validity is one form of assessing validity of the test. In this study, the content validity of the test was analyzed by using checklists to judge the validity of the test. The analyses are presented separately under three sections of the test: reading, grammar and vocabulary, and writing.

Analysis for reading section

The analysis of reading section was recorded in the table as follows:

Table 3: Results of assessing content validity for reading section

	Multiple choice	True/false	Short answer question
Tested getting main idea	-	1	1
Tested skimming skill	-	4	2
Tested scanning skill	4		
Giving opinions	-	-	1
Others	1	-	1
Kind of testing	Direct		
Number of items	5	5	5
Type of scoring	OBJECTIVE	OBJECTIVE	SUBJECTIVE
Answer key is provided	Yes	Yes	Yes
Rating scale is provided	Not Applicable		No

Skills tested

According to the interview data, the reading section was intended to assess students' ability in sub-skills of reading such as getting the main idea, skimming, scanning, and giving opinions on the reading passage.

In the real test, all the above sub-skills are tested. Most of the items tested skimming and scanning skills. Some items tested getting main idea of the reading passage and giving opinions on the passage.

However, the multiple-choice items tested another sub-skill of reading which was not mentioned in the interview. The multiple-choice items tested inference of meaning from context. Furthermore, the short answer questions also tested more than reading ability, as they require the students to produce language in written form. The students need to use their grammar knowledge, vocabulary and prior knowledge to answer the questions. Even though the skill tested is reading, these items require more than using the given information from the reading passage. For example, one of the short answer questions was:

“Of the six character traits of hero which do you think is the most important? Why?”

The answer for this question would need more information than that given in the passage, and the answer would need creativity and expression of ideas which are not testing reading ability.

Texts and Tasks

The reading passage was chosen from one pre-intermediate exercise book. It is a general topic which can be interesting for the students. The reading passage is about the characteristics of heroes. This is not an authentic reading passage. Only one reading passage is given for reading section. There are three formats in the reading section; multiple choice, true/false and short answer questions and each format included five items each. Therefore, there were altogether fifteen items in this section.

Scoring system

The items for multiple-choice and true/false were scored objectively. The responses to short answer questions were scored subjectively. Students get one mark for each correct item for multiple-choice and true/false, whereas they receive two marks for appropriate answers for short answer questions. The scorers took off marks for spelling and grammar errors in the short answers. The short answer questions are supposed to test reading but they are testing much more than just reading skills. There are some difficulties in scoring some short answer questions items, as many answers could be appropriate for the question. As the item mentioned above

“Of the six character traits of hero which do you think is the most important? Why?”

The most important character of hero could be different from one student’s opinion to another. It is difficult to set the standard answer to fit for all students. This question also required the students to give a reason of why they think that specific character is most important. The answer keys for short answer questions are not provided for the scorer.

Analysis of Grammar and Vocabulary section

The analysis of grammar and vocabulary section was recorded in the table as follows:

Table 4: Results of assessing content validity for grammar and vocabulary section

	Gap-filling	Multiple- choice	Matching
Tested present simple tense	1	-	-
Tested present continuous tense	1	-	-
Tested past simple tense	1	-	-
Tested past continuous tense	1	-	-
Tested verb pattern	5	3	-
Tested future intentions	1	2	-
Tested article	5	-	-
Tested preposition	4	-	-
Tested linking words	5	-	-
Tested pre-fix and suffix	6	-	-
Tested word formation	4	-	-
Tested collocation	-	-	5
Others	1		-
Number of items	35	5	5
Kind of testing	Indirect		
Type of scoring	OBJECTIVE		
Answer key is provided	Yes	Yes	Yes

Skills and structures tested

The grammar and vocabulary section was intended to assess the students' use of tenses (present simple tense, present continuous tense, past simple tense, past continuous tense), prepositions, verb pattern, articles, verb and noun that goes together, linking words, word formation, and pre-fixes and suffixes from Unit One to Five of the *New Headway* course book and supplementary vocabulary book. This section uses indirect testing to assess grammar and vocabulary. The test evaluated had all the above structures tested. The results showed that the test content covered all units and the supplementary book. However, only some chosen parts of the structures from each unit were tested in the test in spite of having all the representative samples of the contents from the textbook. Some structures that were taught to the students from the textbook were not included in the test. For example, in *New Headway* unit 4, students have been taught about expressing quantity, using words like 'much and many', 'some and any', 'a few', 'a little' etc. but, those were not included in the test. Moreover, irregular verbs, time expressions, and making negatives from *New Headway* unit 3 were not tested in the test. Besides, only one section from supplementary vocabulary book, linking words, was tested and the content of all the other vocabulary lessons that have been taught from this supplementary book were not included in the test. Therefore, the grammar and vocabulary section covered some part of the structures from *New Headway* unit 1 and to 5 and one section from the supplementary vocabulary book.

Tasks

The test techniques used in this section are gap filling, multiple choice and matching.

This section had eight formats in total of which six were in gap-filling form. The bias for gap-filling techniques is obvious. Among the eight formats, seven formats had five items each and one format had ten items. Therefore, there were altogether forty-five items in this section, twenty-five items were for testing grammar structure, and twenty items were for assessing the vocabulary. The whole section used indirect testing to assess grammar and vocabulary. Indirect testing only measures the abilities that underlie the skills in which the test is intended to assess. Learners are familiar with indirect testing format like “gap-fill” as they have common experiences of being able to complete gap-fills items, but they do not have a chance to use the structure in communication. Therefore, the problem with indirect testing is that we cannot prove the relationship between the performance of the skills we are interested in measuring and the performance in the test.

Scoring system

All the items from this section are scored objectively. Key words are provided and the test takers will get one point for each corrected item. Only one key word for each item is provided and no other possible responses are accepted.

Analysis of writing section

The analysis of writing section was recorded in the table as follows:

Table 5: Results of assessing content validity for writing section

	Essay writing
Tested developing ideas and supporting	Yes
Tested form (organization)	Yes
Tested choice of words skill	Yes
Tested language use (structure use)	Yes
Number of item	1
Kind of testing	direct
No bias in topic	Yes
Length of expected response	150W
Type of scoring	Subjective
Provided specific Scoring rubric	No

Skills tested

According to TD-A, the writing section of the test is intended to assess some sub-skills of writing such as developing ideas, using different structures and choosing appropriate words, organization of the writing, supporting main ideas and being relevant with the topic. The test used direct testing to assess writing ability of the students. The test has face validity as it is testing writing by making the test takers produce a piece of writing.

Texts and tasks

The topic of the essay was “The advantages and disadvantages of living in a foreign country” with no further guidance being given other than the suggestion for number of words. This topic was chosen to cover the writing topics in *New Headway* unit One to Five. There was only one item in this section and the students were asked to produce the writing in essay writing type. The timing for this section was one hour.

Scoring system

This section required subjective scoring. According to TD-A, the essays are marked according to being relevant to the topic, use of sentence structures and appropriate vocabulary, presenting relevant ideas and grammar and spelling. TD-A had considered how to score the writings but no specific marking scheme was developed. TD-A said

“I do not have a rating scale for this. But, I will see whether it is relevant or not, organized or not, and grammar. About spelling, it is not so serious. However, if I compare the essay with lots of spelling errors and only a few errors, I will give more marks to the one with fewer errors. The idea is freethinking but they have to support their ideas well. They have to present their ideas clearly. It is 15% out of 100. I will tell other scorers to look for these points.”

Reliability of the test

The results from the observation and checklists are discussed under two sub headings; the test factors and the scoring factors.

Test factors

The analysis for test factors will be presented in three separate sections; reading, grammar and vocabulary, and writing. The test factors are discussed under two sub headings: (1). test techniques and (2). instructions and test items.

Analysis for reading section

The analysis of reading section was recorded in the table as follows:

PAYYAP UNIVERSITY

Table 6: Results of assessing reliability for reading section

	Multiple choice	True/false	Short answer Question
Number of items	5	5	5
Kind of testing	Direct		
Instruction is clear	No	No	Yes
Language of question is easy to understand	Yes	Yes	No
Length of response	Not Applicable		Not Mentioned
Type of scoring	OBJECTIVE	OBJECTIVE	SUBJECTIVE
Answer key is provided	Yes	Yes	Yes
Rating scale is provided	Not Applicable		No
Only one possible response	5	5	-
More than one possible response	-	-	5

1. Test techniques

Two techniques used in this section (multiple-choice items, true/false) can lower the reliability of the test. There are some problems with using multiple-choices and true/false items; for example, a student can get marks just by guessing, or answer by just writing true for all items or false for all items instead of trying to answer the question thoughtfully. For example, in true/false items:

B. Check the statements that are true.

----- 1. All heroes possess courage, strength, and honesty.

Students can get the marks by just guessing without understanding the information given in the passage. They may have prior knowledge about heroes in their mind and they can get the answer using this.

2. Instructions and test items

For reading section, although the test developers said that the instructions are clear, the instructions for multiple choice items and true/false items are complicated and the students can get confused about how to respond to the questions. For example, the instruction for the 'True/false' item is "*Check the statements that are true.*"

The instruction does not clearly state how the student should respond. This instruction does not mention what to do after checking the statements that are true, or how to respond to the statements that are not true. Therefore, students can come up with many forms of response like writing "true" for the statements that are true and "false" for the statements that are wrong; writing just "T" for true statement and "F" for false statement; or writing just "T" or "True" for the true statement and writing nothing for false statements.

Instructions for the short answer questions are clear but here the language of the questions is complicated and ambiguous. Some students may not be able to understand the meaning of the questions and this will make them lose points. If the language of questions was easier to understand, the results may be different.

For example, one of the ‘true/false’ questions is worded:

“Children need to have heroes.”

This statement is ambiguous as it is not easy to understand what the word “need” means.

The word “need” in this statement can vary from one students’ opinion to another. No definite answer is provided in the reading passage. The answer could be “true” as well as it could be “false”. This question should have clarity.

Moreover, there is no limit for response of the short answer questions. Responses of students can vary one from another and it could reduce the reliability of the results.

Analysis of Grammar and vocabulary section

The analysis of grammar and vocabulary section was recorded in the table as follows:

Table 7: Results of assessing reliability for grammar and vocabulary section

	Gap-fill	Multiple-choice	Matching
Number of items	35	5	5
Kind of testing	Indirect		
Instruction is clear	Yes(15)/No(20)	No(5)	Yes(5)
Language of question is easy to understand	Yes	Yes	Yes
Type of scoring	OBJECTIVE		
Answer key is provided	Yes	Yes	Yes
Only one possible response	27	5	5
More than one possible response	8	-	-

Test techniques

For grammar and vocabulary section, the techniques used here are gap-fillings, matching and multiple-choice. There can be some flaws in using multiple choice items as the students can get marks by just guessing or choosing all '(a)' or '(b)' or '(c)'. However, the techniques used in this section are familiar to students and this can reduce unreliability.

2. Instructions and Test items

The instruction for some of the formats are clear but the instruction for assessing the use of verb pattern and future intentions, verb-pattern, linking words, prefix and suffix, and word formation are not clear. For example, for assessing use of verb-pattern and future intention, the instruction and tasks are illustrated below;

III. Choose the correct form of the verb.

1. *Congratulations! I hear you _____ married.*

1.....

(a) will get (b) are going to get (c) gets

The instruction asked the student to choose the correct form of the verb but did not mention what to do after they have chosen the correct answer. The response format of

students can respond in more than one way, like writing '(b)' in the given space, or writing '(b) are going to get' in the given space.

As some of the instructions are not clear, it can make the students confused about how to respond and they can lose points by not being able to answer in the way expected by the scorer. Moreover, if the instruction is not clear, and restricted to what it intended to assess, it can lead to many possible answer for the students. This following excerpt is taken from the format for assessing tenses.

Complete these sentences using the correct form of the verbs in brackets.

1. I (meet) you at ten o'clock tomorrow. I.....

As the instruction is not stated clearly, what kind of tenses should be use in this format; there can be more than one possible response for this item. This can be "will meet" or "am going to meet" or "am meeting". All the responses are appropriate for the questions, even though the expected answer for this question is "will meet" and other possible answer are not accepted. However, the language of questions of all items is easy to understand.

Analysis of writing section

The analysis of writing is recorded in the table as follows:

Table 7: Results of assessing reliability for writing section

	Essay
Number of item	1
Kind of testing	direct
No bias in topic	Yes
Instruction is clear	Yes
Language of question is easy to understand	Yes
Length of expected response	150W
Type of scoring	Subjective
Rating scale is provided	No

1. Test techniques

The technique used for assessing writing is essay writing. There was only one essay topic and there was no choice. This is good for enhancing reliability of the test as there is only one choice and the responses of the students are limited. The length of the expected response was one hundred and fifty words.

2. Instructions and test items

The instruction for writing question is short and clear and the expected response is limited. The topic is easy to understand. The instruction for essay writing is

“Write an essay on “The advantages and disadvantages of living in another country.”

Write about (150) words.”

However, there are some possibilities that the students will interpret the task in different ways as so little guidance is given. The more differently the task taker interprets the task, the more difficult it will be to compare scores.

The Scoring factors

Scoring the test paper is one of the important parts in testing. The scoring procedure of this test should be explained briefly before discussing the factors that lower the reliability of this test. After brief explanation about the scoring procedure, the scoring factors can be discussed under two sub-headings; 1. scoring system and 2. scorer reliability.

Scoring procedure

The scoring of the answer sheets took place on 10th March, 2007. There were altogether eleven scorers for marking the students' answer papers. Six teachers scored the grammar and vocabulary section (three teachers marked the grammar part and the other three teachers marked vocabulary sections). Two teachers marked the writing samples and three teachers marked the reading answers. However, each item was scored by only one scorer and there was no double scoring. Teachers were given key words to mark for the objective

scoring items but no specific marking scheme was provided for the subjective scoring answers.

1. Scoring System

Reading Section

Both objective and subjective scorings are used in this section. The student will get one mark for each correct item from true/ false and multiple choice and they will get two marks for every appropriate answer in short answer questions. There is only one possible response for all the items from true/false and multiple-choice items and these items are scored objectively. However, it is clear that there is more than one possible response for the short answer questions. The items seem to have many possible answers which would be acceptable. For example, one short answer question is

“Give a suitable title for this passage.”

There can be many possible answers for this kind of question. The test takers could come up with many appropriate titles for the passage and there can be no definite answer key for this kind of question. The examiners scored the short answer questions items subjectively. The problem with the short answer questions is that there is no rating scale for scoring the responses. The scoring is totally depending on the scorer’s opinion. This fact reduces the reliability of the score of the test results.

Grammar and vocabulary

Objective scoring was used in this section. The student will get one mark for each correct item from all formats. Most of the items have only one possible response but eight items have more than one possible response. For example, one item which assesses word formation is as below:

VI. Complete the sentences with suitable words.

“The food is impossible to eat. It is _____.”

There can be many possible answers for this question which are suitable for the sentence. The instruction mentioned that students could use any suitable words for completing a sentence. For those kinds of items, the teachers did not accept the other possible response while marking the paper. The expected answer for above question is “inedible” and the other answers are not accepted. Thus, it can affect the reliability of the students’ results.

Writing Section

Subjective scoring was used for assessing writing ability. The total marks given for the essay is 15 marks and if the essay is readable, well-organized, and free of error, the essay can get 10 to 12 marks, and if it is readable but not well organized with some errors, it is 7 to 9. If the essay is lower than above qualities, the students will get 6 and under. TD-A

gave some points to check for but the scoring of the answers differs from one scorer to another. The scoring became flexible and totally depended on how the scorer judged the writing, as no specific rating scale was provided. Moreover, some scorers counted spelling and grammar errors in essays but some scorers ignored the spelling and grammar errors. Therefore, the subjective scoring of this essay section cannot be claimed as highly reliable.

Scorer Reliability

The scorers of this test are the teachers and the teaching assistant for the first year English course. The experiences of these teachers are range between 30 years to less than one year. They were not trained for scoring the answer papers before the test was administered. Teaching assistants have less than one-year experience of teaching and this is the first time for scoring papers for most of them.

For objective scoring, there is not much problem for scorers, as these items do not need judgment for scoring. However, for subjective scoring, the judgments of the scorers are needed. There are no specific rating scales for either short answer questions or essay writing. The judgments of the scorers may vary from one to another. For example, one scorer might think an essay is good and give high marks while another scorer might not think that the essay deserved high marks and so give lower marks.

Besides, according to the test paper, the students are required to write their name and student ID on the test paper. Having test taker's name on the test paper can effect how

scorers judge individual students and may lead to unfairness for some students and lower reliability in the scoring.

There is no double scoring for the subjectively scored items. This may not be fair for the test takers, as even for the same level written response, the score they received might be different.

PAYYAP UNIVERSITY

Summary of the Chapter

This chapter has presented the results of the study. The first part presented the data on the test construction process of BARS program. The test was constructed by three English teachers and the items were written individually. The test used both direct and indirect testing items and most of the items are objectively scored items. The objectives of the test differ slightly from one test developer to another. The test does not have test specifications. The test items were not moderated or trialled before administering the test. There is no validation of the test nor training of the examiners before the test is administered.

For content validity, the results showed that the test seemed to test what it intended to test in most areas of the test. The test tried to cover the syllabus and included different skills and structures from the textbook. However, some structures from the textbook are not included in the test. Gap-filling technique is highly favored in this test. Most of the test items used indirect testing. In addition, some of the test items tested more skills and abilities than what they intended to assess.

For assessing reliability of the test factors and scoring factors, the findings showed that the test had some ambiguous items and some instructions are not clear. Some testing techniques are not reliable as test takers can get marks by just guessing. Moreover, there is no specific rating scale for subjective scoring and some of the objective scoring items have more than one possible response which is not accepted by the scorer. In addition, the scorers were not trained before administering the test and some scorers did not have experience in scoring examination papers. Therefore, these factors may lead to unreliability.