

CHAPTER V

DISCUSSIONS AND CONCLUSIONS

This chapter presents the summary of the research results, discussions and suggestions for further research.

Summary of Research Results

The results can be summarized by referenced to the research questions.

1. What kind of process is involved in the construction of the test?

The achievement test of BARS program had been constructed by three first year English teachers and the test included three sections: reading, grammar and vocabulary, and writing. This achievement test was based directly on the syllabus.

There were some weaknesses in the test construction process. The objectives of the test differed from one teacher to another and no specific objectives of the test were set before constructing the test. The test did not have specific test specifications but the test developers had general ideas for test contents. Each test developer took responsibility for developing separate parts of the test, made decisions on test content and wrote test items by themselves. After item writing, there was no moderating, trialling or validation of the test before administering it.

The differences between the ideal test construction stages and the actual test construction stages of the BARS program are provided in figure 2.

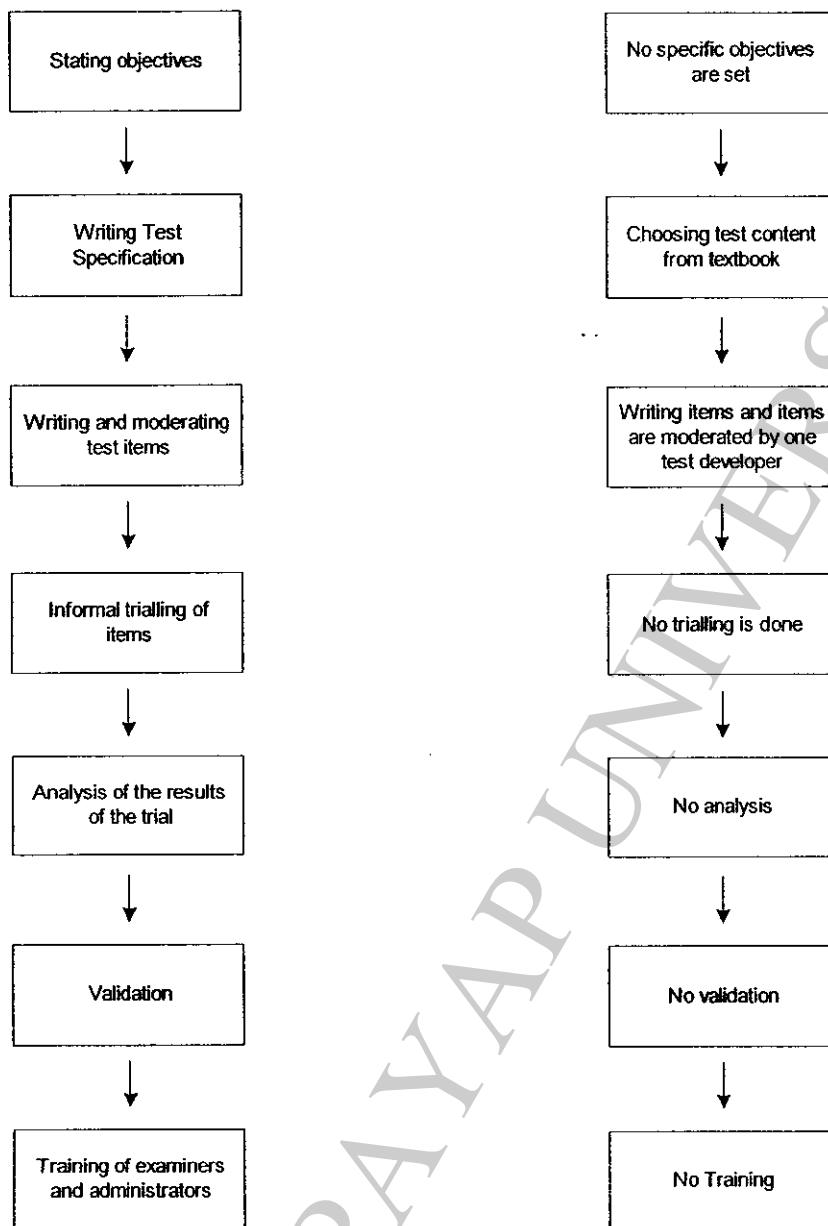


Figure 2: Comparison of ideal test construction stages and BARS test construction Stages

The results of the study showed that test construction does not follow the seven recommended stages of test construction process discussed in chapter II. As the test does not have specific objectives nor test specifications, it is not easy to determine whether the test has achieved its aims or met the objectives of the programs (Hughes, 2003, Alderson et al, 1995). The test may not be a fair reflection of the syllabus and this can make the test unfair for some students. As the test did not have specifications and the test content was chosen by the test developers, the test content seemed to become predictable, which could lead to negative backwash (Hughes, 2003). Apparently, teachers gave hints about what might be included in the test to students. Then the students are likely to just study for the exam and this can lead to the teacher just teaching them for the exam.

The test items were written by individual test developers and the test items were not reviewed systematically after they were written. Moreover, moderation and trialling of the test items are important as it can help the test developers to find the weaknesses of their items and make changes if it is necessary. If the test items are not reviewed after they are written, we cannot be sure whether the test items are appropriate and going to work well for the students or not (Hughes, 2003, Alderson et al, 1995).

However, some test developers thought that their test items were good enough and did not need reviewing. The test developers seem to have different understanding of how important moderating and trialling of items are in developing a good test. They were not of the view that the test needs to have specific test specifications, and that the test should be validated before administering it. Test developers designed the test by their own experiences. It may be that the test developers need more knowledge about language testing in order to produce a better test for the program. Moreover, the program also

needs to arrange language teacher educational workshops and training to improve the skills of the teachers.

How valid is the content of this test for the purpose of assessing achievement?

The test content matched with the test specifications described by the test developers. The test content covered the skills and structures from the test specifications drawn up by the researcher. Most parts of the test seemed to assess what was intended to be assessed. The results showed that the test seemed to have content validity as the test content matched with the content from the test specification (Hughes, 2003).

However, if we look more closely at the validity of the test for assessing achievement, the test is not as valid as it seemed to be. This test assessed only some skills and structures from the syllabus. This test assessed reading, writing, and some vocabulary and grammar structures from the syllabus. Speaking skills, listening skills and some vocabulary lessons from the syllabus are not tested. Some of the grammar structures from the textbook were not included in the test. The test only tested limited skills and structures and could only cover the syllabus to some extent. As this test is an achievement test, the test should include all four skills and structures from the syllabus in order to achieve its objectives or to cover the syllabus (Hughes, 2003; Heaton, 1988; Weir, 1993). In addition, most of the test items are indirect test items. Indirect testing items assess the abilities that underlie the skills in which we are interested (Hughes, 2003). Indirect items may not enhance the validity of the test as the relationship between the performance of the skills we are interested in measuring and the performance in the test are not verified.

These items may result in negative backwash, and the results may only provide limited information about the students' ability. These items may lead to testing on content rather than basing tests on the objectives. Tests should use direct testing in order to enhance the validity and reliability of the test if possible (Hughes, 2003; Weir, 1993; Heaton, 1990).

Moreover, some of the test items tested more skills than what they intended to assess. These kinds of test items are not valid, as they require more of the candidate than the intended skills. Weir (1993) and Hughes (2003) stated that the test should limit itself to measuring only what it is intended to test.

The scoring of some items lack validity as the judging of the items counted more than the intended skills of the students, for example, the scorer counted the grammar and spelling errors in assessing reading. The scorer should score only what the test is intended to measure (Hughes, 2003; Bachman, 1990).

Therefore, there are some limitations with the validity of the test for assessing student's achievement.

3. How reliable is the achievement test being used in this program?

This study was not able to determine the reliability coefficient of the test as the test structure was not appropriate for applying the split-half method for estimating the reliability of the test. However, the findings for this research showed that there are some test factors and scoring factors that are likely to lower the reliability of the test and the scores of the students.

For test factors, some test techniques (for example; multiple-choice and true/false) used in this test could be answered by just blind guessing and this may effect on the reliability of the score. Moreover, the test had some ambiguous items and had unclear instructions which can confuse the students. In order to enhance the reliability of the test, the test instructions and items should be clear and unambiguous for all the test takers (Davies & Pearse, 2000).

Furthermore, the scoring of the test cannot have high reliability, as there are some flaws in producing and scoring of the objective scoring items and subjective items. Almost all of the formats (except two formats) were designed to be scored objectively. Therefore, this test should be highly reliable (Bachman, 1990). Hughes (2003) mentioned that objective tests could have high scorer reliability, as they do not require any form of judgment from the scorer. However, some test items have more than one possible response but the scorers accepted some of the possible responses but not all. Therefore, some students lose points even though they offered possible responses that should be acceptable.

There is no specific rating scale for subjective scoring and the judgments for subjective scoring vary from one examiner to another. The subjective scoring requires a specific rating scale in order to increase the reliability of the test (Weir, 1993) but this test did not have rating scale for subjectively scored items. Besides, each answer paper was scored only once by only one examiner which meant that there was no double scoring for both subjective and objective items. Actually, more than one scorer should score the exam papers and double scoring should be applied in order to reduce unreliability of the scorers (Hughes, 2003).

Additionally, examiners were not trained before administering the test and scoring the exam papers. The judgments of the scorers vary from one teacher to another in scoring items. Teachers should be trained well for the administering of the exam and scoring of the exam papers (Hughes, 2003; Alderson et al, 1995).

To summarize the above points, the findings showed that the test had some weaknesses and the content validity and reliability of the test factors and scoring factors were judged to be not high.

Reasons for weaknesses in the test

There are two main reasons which caused the test to be a rather poorly constructed test.

These reasons are

1. Test developers' lack of awareness of language test construction
2. Limited resources (human, time, materials, financial support)

1. Test developers' lack of awareness of language test construction

Test developers might need to know about some important factors in language testing.

They were not fully convinced about the need to construct a test by reference to procedures that are mentioned in the language testing literature. They had constructed the test without test specifications, test objectives or other essential stages of test construction according the way of constructing test which they are used to. Moreover, the test developers seemed not fully aware of the importance of validity and reliability in the test.

The test developers' different understanding of language test construction led the test to being less valid and less reliable than it might otherwise have been. If the test developers have more knowledge about language testing, the test might have been constructed in a more satisfying way.

2. Limited resources

The other factor that has led to weaknesses in the test is the limited resources.

There are only six teachers in the first year course and only three teachers can give time to construct the test. The limited human resources cause the test to become less reliable as there are not enough teachers to review the test items. Moreover, the scoring of the test may be unreliable as there were not enough scorers to have double scoring. As the program has only some limited experienced teachers and the students' results needed to be handed in on time, teaching assistants were asked to score the test papers. However, teaching assistants are not qualified to administer the test.

Teachers could not give time to work together for construction of the test. The test duration is only three hours and it is not enough time to include all the items from the textbook. Teachers could not give time for administering a speaking test. Speaking and listening tests could not be included as they are time consuming for both teachers and students to construct, administer and take. Also material support such as necessary cassettes, cassette tapes, electricity supplies and other materials are not sufficient to have speaking and listening tests. This therefore is concerned with the limited financial support the test has.

Suggestions for the development of the test

- 1) Speaking and listening skills should be tested as students have learnt these skills during the semester. All skills and structures from the syllabus should be included in the test in order to achieve the objectives of assessing student's achievement. Moreover, all four skills are important for learners. If listening and speaking are not included in the test, some teachers might omit teaching listening and speaking activities. Teachers and students can treat these two skills as unimportant skills. Therefore, the test should include listening and speaking skills. One possible way to assess listening is to include the listening part for about 20 to 30 minutes in the test. If the total time of the test is 3 hours, the test could be divided into 4 parts, 30 minutes for listening, 30 minutes for writing, 1 hour for reading and 1 hour for grammar and vocabulary. The materials for the listening test can be taken from resources like other *Headway* pre-intermediate books, other printed pre-intermediate level listening practice and textbooks, and online listening practice resources. Alternatively, the test developers can take some excerpts from resources like news, native speakers' dialogues and movies to develop in-house listening tests. For assessing speaking, there will be some difficulties to assess speaking skills of over 140 first year students. However, the assessment of speaking can be done as in-class tutorials during the course instead of including it in final achievement test. As *New Headway* provides speaking skills practice in each unit, the teachers can assess their students' speaking skill at the end of every unit by using role-play, discussions, individual presentations and group projects.

- 2) The test should include more realistic language activities to perform under appropriate conditions. The test could include texts that are more authentic and tasks that require students to perform the language as in real life situations. For example, the test could include listening to excerpts of real conversation between native speakers and require the students to listen and answer questions. The test could use authentic texts like excerpts from magazine articles, newspaper articles and news journals.
- 3) Teachers should be aware of theoretical issues in language teaching in order to develop the quality of the test. Teachers need more awareness of language teaching and language testing. Therefore, the program administrators should provide more training and workshops for the teachers to enhance their language teaching knowledge especially in language testing areas. The training and workshops on designing and constructing tests for specific purposes, writing test items and scoring of the exam papers can be conducted for the teachers and test developers.
- 4) Tests should be designed systematically by setting specific objectives, drawing test specifications and following the test construction process carefully. Test contents and items should be selected carefully in order to cover the syllabus. Test items should be reviewed and checked by other teachers after they are written. Tests should have specific rating scale for subjective scoring if they include the

subjectively scored items. Test layout should be appropriate and the test should not have any spelling mistakes or typing error. Moreover, for the reason of anonymity students should only be required to write their students ID number.

- 5) When developing a test, all the teachers from first year classes should work together in order to design appropriate course objectives for first year courses. They should work together from setting objectives up to scoring of the items. All teachers must take responsibility in designing and constructing the test. The program administrators and the head of the department of English should instruct the teachers to work together or assign the duty for the teachers and help them in producing the test.
- 6) Students should not be given hints on what is going to be in the test. The test should be fair for all students. Students should be encouraged to study all the lessons they have learnt. As the aim is assessing students' achievement over one semester, the teacher should not make them study only some parts of the textbook for the test. This will not help in assessing student achievements.
- 7) Teachers should be trained for administering and scoring of the test. Examiners should be chosen carefully and only qualified teachers should administer and mark the test. Before marking the answer papers, the scorers should be oriented in detail about the scoring procedure and the rating scale. Inexperienced teachers should be allowed to score the paper only after receiving appropriate training.

Checklist for construction of an achievement test

The construction of an achievement test is an important process for test developers and programs. There are some points that need to be considered when constructing tests. This section provides a checklist for constructing an achievement test for test developers and programs. This checklist contains eight different categories for use before and after constructing a test.

PAYYAP UNIVERSITY

Table 9: A checklist for constructing an achievement test

Objective	<ul style="list-style-type: none"> ○ Does the test have specific objective/ objectives?
Test specifications	<ul style="list-style-type: none"> ○ Does the test have test specifications? ○ Do the test specifications provide detailed information on test content, test structure, timing, medium channel, techniques and scoring procedure?
Test content and items	<ul style="list-style-type: none"> ○ Does the test assess the full range of appropriate skills and structures, as defined by the test specifications? ○ Does the test use direct testing items? ○ Are the test formats familiar to students? ○ Does the test use authentic text?
Scoring procedure	<ul style="list-style-type: none"> ○ Are answer keys provided for the objectively scored items? ○ Does the mark scheme anticipate responses of a kind that candidates are likely to make? ○ Is a rating scale provided for the subjectively scored items? ○ If so, can the rating scale be easily interpreted by a number of different examiners in a way which will ensure that all mark to the same standard? ○ Are the scorers experienced in scoring exam papers? ○ Is double scoring provided?
Layout	<ul style="list-style-type: none"> ○ Is the layout candidate friendly? Clearly typed and printed? ○ Does the test format provide enough space for responses? ○ Does the test free of any kind of error? (grammar, spelling, typing) ○ Is the test trialled?
Trialling	<ul style="list-style-type: none"> ○ Can the tasks be answered satisfactorily in the time allowed? ○ Does the test set an appropriate level of difficulty? ○ Are there any difficulties in administration of the test? ○ Are there any difficulties in scoring of the items?
Analysis	<ul style="list-style-type: none"> ○ Are the tests items edited and revised by the test developers after trailing?
Training	<ul style="list-style-type: none"> ○ Does the program provide a handbook, which contains the rationale of the test, description of the test, an explanation of how test scores are to be interpreted and details of test administration? ○ Are the examiners trained before administering the test?

Limitations of Study

This research included both quantitative and qualitative data. There were some limitations about the interview data, as some participants were not eager to respond freely as they did not want their test to appear to be of poor quality. Moreover, this study did not include comments and opinions from the students who took the exam. Besides, this is only a case study on one first year achievement test, and the results of this study could only evaluate the content validity of test content but not other ways of establishing validity.

Further research

This study was conducted to evaluate the achievement test of a particular language program. Based on the results and suggestions for development of the test, an action research project on implementing the suggestions here could be carried out for this BARS program. This action research could give insights into how far the suggested factors can work for the improvement of the test quality and what kind of problems might be faced when putting theory into practice.

Further research could include studies carried out for evaluating the placement test used in this program, with the aim of improving the quality of the placement test.

Besides, this study only evaluated a first year achievement test. Further research studies could be conducted for evaluating the whole assessment system of this program,

which will include first year to final year. This may be beneficial for the program to see the strengths and weaknesses of their assessment system.

Moreover, other aspects of both internal and external validity such as construct validity, predictive validity, and concurrent validity could be investigated for further research. The results of the test could be correlated to find out the construct validity and concurrent validity of the test.

PAYYAP UNIVERSITY