This chapter presents a summary of the study, its findings and a discussion of these findings. It also presents the recommendations of the study for test writers, test developers and test users in Yemen and in similar contexts of admission testing. The chapter closes with the limitations of the study and its suggestions for further research.

Summary

This study was carried out to achieve two purposes. The first purpose was to evaluate the 2005 in-house admission test used at Taiz University. The study reached the conclusion that this test was not a sufficient measure of the test takers’ language ability and that the test itself and its administration procedures needed to be revised. The researcher reached this conclusion using two types of evidence: quantitative evidence and qualitative evidence. The type of quantitative study or statistical analysis used in this study was dictated by the type of information available. The only information available to the researcher was the number of test takers and their scores – their total scores, their scores on each test question, and on each test item. Based on this information, the research conducted three types of statistical analysis: descriptive statistics, classical item analysis and ‘Cronbach’s Alpha if Item Deleted’. On the other hand, the qualitative evidence was gained from analyzing three questionnaires given to the test writer, test scores and raters, and test takers. The statistical analysis of the test scores and the feedback gained from the three qualitative research instruments gave the researcher a clear picture of the weaknesses of the test and of the shortcomings of the admission testing situation in general.

These weaknesses and shortcomings guided the researcher in achieving the second purpose of this study, i.e. constructing a more effective and valid admission test. This test was based on one of the most comprehensive models of communicative competence – Bachman and Palmer (1996). The Bachman and Palmer (1996) model was chosen because of its amenability to any kind of statistical analysis, especially for validation purposes. Besides, all the steps of test development in the model are clearly specified, which enables any researcher in the field to make a comprehensive analysis of the test in particular and of the whole testing procedures in general. Realizing the
comprehensiveness and wide applicability of this model, the research tried to follow all the steps in test development starting from writing the test specifications, through writing the test items to the reporting and archiving of the test. The model was, however, subject to slight modifications in order for it to meet the particular characteristics of the research population and of the research situation. In adopting the Bachman and Palmer model in the test construction part of the study, the researcher’s aim was to ascertain the applicability of the model to the context of admission testing in Yemeni universities. The proposed test was intended to be fairer, more comprehensive, and more professional. The test was also eclectic. It measured linguistic as well as communicative competence using a combination of discrete point and integrative tasks. This was achieved by considering, as far as practically possible, the minimum acceptable levels of the six qualities of usefulness and by taking into account the limited amount of the resources available at Taiz University. This test was also intended to meet the prerequisite level required for admission into the English courses at Taiz University and the minimum acceptable level of proficiency of the test takers. Finally, the primary purpose of the study (in its evaluation and test construction parts) was to initiate a much needed, and long overdue, professional admission testing program at Taiz University by creating a set of practical procedures which could be used in the production and evaluation of subsequent in-house English admission tests.

**Findings and Discussion of Findings**

Based on the results of the study, the following findings were recorded. The findings are divided into two sections: findings gained from the statistical analysis (the three types of statistical analysis), and findings gained from the qualitative study (the questionnaires administered to the test writer, the test takers, and the test scores and raters). These two sections are discussed below.

**Findings from the quantitative study**

1. The descriptive statistics, particularly the mean values and the positively skewed distributions of questions B, E, G, H, and I, indicated that these questions were difficult for the majority of test takers, while the mean values and the negatively skewed distributions of questions C and F indicated that these two questions were very easy for all the test takers. This finding was supported by the IA analysis which showed that
questions B, E, G, H and I contained a large number of difficult items and no easy items and that questions C and F contained easy items with no difficult items. This clearly shows that the difficulty level of the questions was not considered while developing test items and, therefore, the different ability levels of the test takers were not fairly considered while developing the items for each question and consequently for the test as a whole. This must have negatively affected the selection of the best test takers.

2. The results of the IA analysis showed that out of the 76 items of the test, 34 items were difficult (their IF values were below 0.29) and only 9 items were easy (their IF values were above 0.71). This means that the majority of test items were difficult for all the test takers.

3. The IA analysis indicated that 34 of the 76 items were also poor discriminators – their ID values were less than 0.29. This means that these items did not discriminate among the high level test takers on the one hand and among the low level tests takers on the other. All test takers could answer the easy items and only the test takers with high level of language ability could answer the difficult items. There were no test items to discriminate between test takers whose language ability falls in between the high and the low levels of language ability.

4. The number of the items in each question was very small, which negatively affected the reliability of the questions. Questions C, D, F, H and I consisted of a maximum of seven items, which is generally not enough to form a reliable scale. Furthermore, after conducting the IA analysis and identifying the items that should be eliminated, 33 items were eliminated. These were the items which were unacceptable in terms of both their IF and ID values. Consequently, the alpha reliability was reduced even further (since alpha is reduced if the number of the items is also reduced). The ‘Alpha if deleted’ analysis indicated clearly how alpha was reduced when these items were deleted. The ‘Alpha if deleted’ analysis also identified other problematic items (item 4 from question F, item 6 from question E, and item 4 from question D) that should be deleted in order to increase the reliability of the questions.

5. The rating of questions J and K was not based on any rating scale. They were also not rated twice by one rater or by two raters. These drawbacks of
the rating of these two questions did not make it possible to estimate the reliability of these two questions (whether intra-rater or inter-rater reliability or the reliability of the subscales).

6. The results of analyzing the test as a whole and of analyzing the questions separately showed that the problems with the test as a whole were less serious than those with the individual questions of the test. This finding about the test as a whole is explained by the large size of the sample (969 test takers) and the large number of the items. The test contained 76 items and two subjective questions, compared to the small number of items in the questions (which at times was as small as four items in a question). High-stakes language proficiency tests should contain a large number of items, and so should the different sections of these tests.

**Findings from the qualitative study**

1. More than 66% of the students reported that the test concentrated on some skills and ignored others.

2. There were two obstacles to the test takers’ best performance. The first obstacle, according to over 70% of the test takers, was the insufficient time of administration. The second obstacle, according to over 67% of the test takers, was the unhelpfulness of the proctors. These two elements combined to negatively affect the test takers’ test performance.

3. About 60% of the test takers considered the test “completely unrelated” to the first-year syllabus of the departments. This indicates that the TLU domain was not considered in writing test tasks and test items.

4. More than half the sampled test takers (58.70%) were dissatisfied with the score distribution of the whole test. Over 72% of the students reported that grammar was given more space and more scores. This finding is borne out by the researcher’s review of the test (Table 3.1). Of the total 76 test items, 34 items tested knowledge of grammar. The scores seem to have been assigned according to the number of items on the test tasks and not according to the importance of the language ability tested.

5. About 35% of the sampled test takers were not satisfied with their performance on the test. The test takers attributed their dissatisfaction to three reasons: 1) the difficulty and the unexpectedness of the test tasks (56.25% of the respondents), 2) the unclear instructions and the noisy
examination halls (25% of the respondents), and 3) the insufficient time of administration (18.75% of the respondents). This clearly indicates that the administration procedures did not help the test takers give their best performance, which led 91.30% of them to believe that if they were given a chance to take the test again, they would do better.

6. The test writer’s concern with ‘merit’ is not measured with relevance to a particular level of language proficiency but rather with relevance to the overall performance of the test takers on the test. A student need not achieve any specified level of competence in order to qualify for admission but needs to do better than the other test takers. In this case, the specification of a pass mark (or cut-off score) becomes meaningless and the purpose of the test changes from being concerned with selecting the ‘best’ test takers to being concerned with selecting the required number of test takers regardless of their proficiency.

7. The test writer assigned the test scores according to “the difficulty of the task” and not according to the “importance of the language ability tested”. The precedence of task difficulty over the importance of language ability raises a number of questions. Is the difficulty of the task a more dependable criterion than the importance of the language ability tested? What tasks should be ‘made’ more difficult? And why should a particular task be made more difficult than another? The test takers’ scores on these tasks, and on the test as a whole, were, according to the test writer, “not always the best indicators of the test takers’ level of language proficiency”. Assigning scores according to the difficulty of the task may, therefore, not be the most dependable criterion of assigning scores.

8. Three of the test scorers and raters reported that the scoring team was “not really efficient” and ascribed their dissatisfaction to two reasons: 1) that some of the team members did not have enough experience in test scoring and 2) that some of the team members wanted to finish their job fast because they thought they were not paid well. These problems with test scoring might have affected the fairness of the scoring.

9. There was a surprising disagreement among the four test raters on how essay questions were rated. Two of them reported that they consulted fellow raters to decide on a strategy, the third rater relied on intuition.
whereas the last rater followed the guidelines of the test writer. The absence of agreement among the raters must have made the rating of these questions unfair and the absence of written guidelines or rating scales risked making the rating inconsistent. Besides, the absence of clear formal rating scales made three of the test raters take the performance of a good candidate as a scale against which they rated the responses of the other test takers. The test raters expressed dissatisfaction with the “random rating” caused by the absence of rating scales. This dissatisfaction and ‘randomness’ could have been minimized if rating scales had been made available to the test raters in advance.

10. Twelve of the test scorers and raters reported that their main problem with the scoring was caused by “the nature of the test tasks”. This remark is supported by the researcher’s overview of the test, which showed question G to be problematic and confusing. If the test tasks were confusing to the test scorers, they must have naturally been more confusing to the test takers.

11. Ten of the test scorers and raters referred to the test writer in order to solve problems and resolve differences over the appropriateness of the test scores, while four others accepted the scores as they were even if some of them were found to be randomly or inappropriately assigned. What this indicates is that the assigning of test scores had not been considered carefully before the administration of the test. This could have caused scoring problems and necessitated changing the scores after test administration. In either case, the testing procedures became rather unfair.

12. Four of the test scorers and raters indicated that the ‘test team’ needed training in test administration and scoring and that the admission tests should test language abilities rather than “abilities in test-taking”. Three other scorers and raters recommended that admission tests give due attention to the speaking and listening abilities.

Recommendation to Test Writers, Test Developers and Test Users

The following recommendations for test writers, test developers and test users were prepared on the basis of the findings of the study. They aim to give practical advice to test writers at Taiz University on the processes of test construction, to test
developers on the procedures of test development and administration, and to test users on test use. The recommendations concerning test construction were followed by the researcher in the construction of the test in the fourth chapter and should be followed in the construction of subsequent admission tests at Taiz University.

1. Admission tests should be constructed on the basis of a clear theoretical model. This would help test developers investigate the reliability of the scores and the validity of the inferences made on the basis of these scores.

2. Admission tests should pass through clear and organized stages of test development and all the processes in each stage should be recorded in order to help the test developers and test writers make any necessary revisions in the test development processes.

3. The test specifications should be specified right from the beginning. These specifications should include a description of the purpose of the test, a description of the test takers, an identification of the resources that should be allocated for the different stages of test development, a description of the TLU domain from which the test tasks will be selected and a clear definition of the constructs to be measured.

4. The different sections and tasks of the test should be described in terms of importance of tasks, sequence of tasks, salience of tasks, number of items in each task, and the scores of each task and each item in the task.

5. Admission tests should be divided into clear parts or subtests under which clear language abilities should be tested using an adequate number of tasks and of items (at least 20 items per part). The scores should be assigned according to the importance of each language ability and according to the number of items used to elicit the required level of that ability. This will enhance the face validity of the tests and the reliability of the test scores.

6. The difficulty level of the items in each section and in each task should be compatible with the different levels of ability of the candidates. Easy items should be used as a lead-in at the beginning of each section and of each task in order to put the test takers at ease and to discriminate at the lowest level of ability. Similarly, difficult items should also be used at the end of each section to discriminate at the highest levels of ability. Moderate items should be used between the easy items and the difficult ones to make the distinction more reliable and the selection more efficient.
7. The pass mark (or cut-off score) should be set at an appropriate level of language ability. Language testing experts with the help of the language teachers should look at the sections or subtests of the test and estimate what score “a barely-adequate candidate” (Alderson, Clapham, & Wall, p. 158) ought to achieve on each subtest. The cut-off score for the whole test should be set accordingly. No test takers should be admitted to the courses unless they reach the level of cut-off scores.

8. After writing the first version of the test, this version should be examined by qualified professional testers and by the teachers who have enough information about the test takers and the testing situation at Taiz University. After revising the first version, the modified version should be tried out. This modified version should be piloted to a similar group of test takers (test takers with similar personal and background characteristics) and their scores should be analyzed statistically. Qualitative feedback should also be collected in order to arrive at the most suitable version of the test.

9. The items eliminated after the IA analysis and the ‘alpha if deleted’ analysis should be either revised or replaced by items with acceptable IF and ID values. The number of items, however, should not be reduced so that the alpha reliability of the subtests and of the test as a whole does not decrease. This practice will enhance the effectiveness of the test before its actual use.

10. After administering the actual test, the test takers’ scores should be analyzed statistically and qualitative feedback should be collected from test takers, test administrators, test scorers and raters, and test users. These two types of information should be collected after each test administration in order to arrive at the best test that suits the test takers and meets the requirements of the English programs at Taiz University.

11. Essay questions should be rated by two different, and well-trained, raters in order to investigate the inter-rater reliability of these questions and make statistical analysis possible. The rating of these questions should be based on clear, sufficient and efficient rating scales which would help test raters rate essay questions systematically and help test developers
investigate the reliability of the subscales for a better way of analyzing such questions.

12. All the information related to the test, its purpose and use should be reported after the test administration. These reports on the test can be used by many people and institutions. They can be used by Taiz University as a testing institution in order to revise and improve the test, by the teachers in order to prepare potential candidates for similar tests, by other testing institutions interested in using the test or by language testing researchers who are interested in evaluating the test.

13. The test items should be stored on a computer along with the information related to each item in the test. This process of archiving the test is important for constructing subsequent tests. The test writers will have available at their disposal a bank of test items (and information related to previous use of these items) out of which they can select the items that are appropriate to their own test purpose and use.

14. Test takers should be informed early and clearly of everything related to the test – its purpose, the abilities it is intending to test, the scoring criteria, the format of the test, the instructions for the whole test and the instructions of each subtest and each task, the time specified for the whole test as well as for each subtest and each task, the description of the method of testing, and the description of all the tasks included in all the subtests. Preferably, this information should be contained in a booklet to be distributed to the test takers at least three months prior to taking the test.

Limitations of the Study

This study has its own limitations, which may be outlined as follows:

1. This study is limited to the context of Yemeni school leavers intending to major in English at Taiz University. The findings of the study are generalized to this particular sample of test takers.

2. The researcher evaluated the 2005 admission test because it was the last admission test administered before she started her research work. The evaluation was restricted to only one test because statistical analysis, and specially item analysis, requires such information as the number of test takers, the total score of each test taker, and the scores of each test taker in
each test task and in each test item. This kind of information was not available for any other test because of the absence of any archiving system at the English departments. The researcher managed to get the information needed to evaluate the 2005 test because it was administered during her presence at Taiz University.

3. Item statistics or item analysis was conducted using classical item analysis (CIA) because of the availability of the software (Microsoft Excel and SPSS) used to conduct this type of statistical analysis. Item response theory (IRT) analysis could not be conducted because of the unavailability of the software used to conduct IRT analysis (e.g. the XCalibre program).

4. The researcher did not conduct a validation study of the 2005 test because the test was not constructed on the basis of a theoretical framework and therefore the information necessary for any validation study was lacking. The information required to conduct a validation study includes formal and well-developed specifications of the test, which must contain a clear description of the purpose of the test, the definition of the construct(s) to be measured, the TLU situation, the allocation of the resources needed, and the characteristics of the test takers.

5. The proposed test was developed by the researcher during her PhD work in India. The researcher could not administer the test while studying since the admission tests at Taiz University must be developed by teaching staff (currently teaching) holding a PhD degree (ranked as an assistant professor or above). After completing her research study and resuming her teaching career, the researcher will be in a position to administer the test.

6. The researcher did not include the try-out of the test in the process of test development for two reasons. First, trying-out the test would have risked making its content known to its actual test takers. Second, the test try-out requires extra resources (in terms of human resources, material resources, and time) which are difficult to obtain at Taiz University unless for an actual test administration.

7. The booklet of test specifications and the post-test questionnaires were not attached to this dissertation for considerations of space. They will, however, be prepared by the researcher and made available in due time.
Suggestions for Further Research

1. A follow-up study should be conducted in order to find out whether or not the proposed test achieves its purpose. This study would ascertain the applicability of the Bachman and Palmer (1996) model and its effectiveness as a model for proficiency test construction in the context of admission testing in Yemen.

2. A follow-up study should also be carried out on the item characteristics, reliability and validity of the four subtests of the proposed test as well as the social consequences and value implications of score interpretation and use.

3. It would be an even more important step for more advanced research to evaluate the proposed test using more sophisticated quantitative methodologies such as generalizability theory (G-theory), item response theory (IRT) and structural equation modeling (SEM).

4. It would also be very interesting for future research work to find a way to include topical knowledge and strategic competence in the construct definition and to evaluate them and collect feedback about them after test administration.