CHAPTER FOUR
THE PROPOSED TEST: STAGES OF TEST DEVELOPMENT

The results of the quantitative and qualitative analyses of the 2005 admission test strongly suggest a need for a more appropriate assessment instrument. This instrument is a test whose rationale should be quite transparent to those involved in its construction, those who have to justify its use and the decisions taken on the basis of the candidates’ scores on it. This chapter presents an attempt to construct one such test.

Since the realization that one should no longer rely solely on a test of linguistic competence, which may be reliable yet lacking in validity, the researcher gives more importance to the different components of language abilities included in the four language skills, i.e. reading, writing, listening and speaking. These four skills are usually not taught in isolation. They are always integrated with each other in natural discourse. The researcher, therefore, attempts to balance the attention given to ‘discrete point’ tasks and ‘integrative’ tasks in order to enhance the reliability, validity and authenticity of the test.

In constructing any high-stakes language test, there must also be practical constraints such as time, money, the imagination of the test constructor and the need for reliability and validity. This will always temper how realistic test events can be made. The researcher wishes to realize the specifications completely in the proposed test in order for the test tasks to reflect a better picture of reality (the TLU domain), taking into consideration these practical constraints on test development. In proposing the test, the researcher follows the theoretical framework proposed by Bachman and Palmer (1996). This theoretical framework provides the scaffolding and the particular features of the Yemeni admission testing context serve to fill in the spaces. The use of theoretical frameworks in test construction is very important because it provides a benchmark against which language tests could be judged and evaluated against the purpose of their development and use. The three stages of test development of the proposed test are specified clearly to enable any language testing researcher to conduct any type of study on the test, whether quantitative or qualitative.
STAGES OF TEST DEVELOPMENT

Bachman and Palmer organize the entire process of language test construction and use into three stages: the design stage, the operationalization stage and the administration stage. In the design stage, Bachman and Palmer describe in detail the different components of the stage, which are listed in a document called a ‘design statement’ (or test specifications). The design statement is intended to ensure that the test takers’ performance on the test tasks correspond closely to the language use tasks. Furthermore, it will also help ensure that the test scores are “maximally useful for their intended purpose” (1996, p. 86). The design statement will be discussed in detail below. In the operationalization stage, the specifications of test tasks that will be included in the test will have been developed. After that a blueprint will be developed to describe how these different types of tasks will be organized to form the actual test. This will lead to developing and writing the actual test tasks. Finally, the test instructions will be written and the procedures for scoring the test will be specified. In the administration stage, the test is given to the test takers, the information regarding test scores and the test takers’ feedback are collected. This information should be analyzed to assess the usefulness of the test and to make the decisions for which the test has been developed.

The following is a detailed description of the use of these stages in developing the proposed test.

Stage 1: Test Design

As pointed out earlier, this stage includes writing a design statement (or test specifications), which describes the six different components of the test design. These six components of the design stage are discussed below.

Description of the Purpose of the Test

This test is designed to assess the English language proficiency of Yemeni secondary school leavers who intend to join undergraduate English courses at the faculties of Arts and Education at Taiz University.

A. Making inferences

The aim of this test is to make inferences, based on the test scores, about the test takers’ ability to understand and produce a type of English that is sufficient to meet the requirements of the future course of study.
B. Making decisions

This test is a high-stakes test. High-stakes decisions have a major impact on the lives of the test takers in particular and on the programs in general. The decisions made on the basis of test takers’ performance in the test help test developers select and admit the test takers who are most likely to succeed in a college-level program. The test takers admitted should be able to 1) understand written English for reading textbooks, articles, short stories and other sources of information, 2) produce acceptable written English for formal academic writing tasks, 3) understand spoken English such as classroom lectures and formal and informal discussions, and 4) produce acceptable spoken English for taking part in formal and informal conversational exchanges of different lengths and different purposes. High-stakes decisions are also important in order to prove the effectiveness of the testing program in achieving the purpose of the test. These selection decisions will be used to determine whether to continue with this new program as it is, or to modify it, or even replace it with a more efficient program.

Description of the Characteristics of Test Takers

There are certain characteristics of the test takers that must be considered when determining the characteristics of the test tasks. For the present research context, these characteristics are:

A. Personal characteristics of test takers

These characteristics include the following:

- **Age**: 19 to 20
- **Sex**: males and females
- **Nationality**: Yemeni
- **Native Language**: Arabic
- **Level and type of general education**: secondary school leavers
- **Type and amount of preparation or prior experience with the test**: all test takers will be familiar with the test format from the booklet that will be prepared and given to them before taking the test.

B. Topical knowledge of test takers

This test will be given to a homogeneous group of test takers who share similar language and cultural backgrounds and who intend to undertake similar courses of study. These test takers share specific topical knowledge such as their
experience of studying English at school and their opinions about the difficulties of learning English. This kind of topical knowledge is shared by test takers and can be used as an information base that helps them demonstrate their language ability. The researcher, therefore, intends to include test tasks whose characteristics are common to the TLU tasks of all the test takers and avoid using tasks that favor test takers who may have high levels of specialized topical knowledge.

C. Levels and profiles of language knowledge of test takers

The general level of ability of the test takers generally ranges between novice high and intermediate low (ACTFL Proficiency Guidelines, 1999). The test tasks will, therefore, be designed for test takers who have language ability ranging between intermediate mid and intermediate high. This level of language ability of the test tasks is the minimum prerequisite level of language ability that will help test takers understand written textbooks, understand classroom lectures, write adequate academic writing tasks and participate in class discussions in an adequate way. Therefore, these test takers should improve their language ability in order for them to reach this prerequisite level in reading, writing, listening and speaking. The test takers have a one-year gap between secondary education and undergraduate education, a long enough time for them to try and reach the level of language proficiency required. Furthermore, the booklet of test format and specifications will have been made available to all test takers. This booklet, in addition to the sufficient time given to the test takers, should help them improve their language abilities in accordance to the levels of the test tasks. The test takers also have the option of joining English classes offered by private institutions in preparation for taking the test.

D. Possible affective responses to taking the test

Test takers’ performance on the test will be affected directly by their personal characteristics, topical knowledge and level of language ability. Test takers whose personal characteristics help them perform well in responding to the test tasks will have positive affective responses to the test tasks. Similarly, test takers whose topical knowledge is relevant to the academic setting will have a positive feeling about taking the test. Finally, test takers with higher level of language ability will feel positive about taking the test. Other test takers may feel threatened by the test because they may not meet the minimum standards for admission into the English programs at Taiz University.
Inventory of Available Resources and Plan for their Allocation and Management

A. Identification of resources that will be available for carrying out test development activities at each stage

1. Design stage

   a. Human resources

      1. Test writer and test developer: the researcher, in this project, will assume the role of the test writer, who writes the test tasks, and the role of the test developer, who supervises the whole process of test development from specifications to administration, use and archiving.

      2. Clerical support: there is no clerical support at this stage. The researcher herself will type, print and photocopy the test materials during the specification and operationalization stages of test development.

   b. Material resources

      1. Space: the researcher’s study room is the place in which test specifications are designed.

      2. Equipment: the equipment needed at this stage is a personal computer and a printer to type and print the test specifications. This equipment is available with the researcher in her house.

      3. Time: developing the test is a part of the researcher’s PhD work. There are no time limits imposed on the researcher.

2. Operationalization Stage.

   This stage requires the same human and material resources required for the design stage. In addition, this stage requires two well trained raters to train the raters of the writing subtest, and two well trained raters-examiners to train the raters-examiners of the speaking subtest in examining the test takers and rating their responses. Two separate, well-equipped rooms and sufficient time will be made available to these raters and examiners.

3. Administration and use stage.

   Since this test will be administered and used after completing this research work, the researcher and her equipment will not be the only resources. Some other human and material resources will be required.

   a. Human resources

      1. Test developer: the researcher will be the test developer in order to ensure that all the procedures of test administration and use are strictly followed.
2. Clerical support: a) one department secretary who will photocopy the test materials two days before administering the test, and b) one department administrative assistant who will keep records of test takers’ scores.

3. Test administrators or proctors: should be teaching staff members of the English departments at Taiz University. Since the longer part of the test is scripted, little training may be enough to administer the test. In the case of the speaking subtest, the examiners should be teaching staff who have had enough training in interacting with test takers.

4. Scorers/raters: scorers for the reading and listening subtests and raters for the writing and speaking subtests should be well-trained teaching staff currently teaching English courses at Taiz University.

b. Material resources

1. Space: an office should be provided for the test developer and the other teaching staff involved in the administration and use stage.

2. Equipment: the equipment needed at this stage is an office computer, a printer and a copy machine to type and photocopy the test.

3. Time: no released time is provided. Test development is considered part of normal work load for test developer and writer.

4. Test materials: booklets, pencils with erasers, a pencil sharpener, individual chairs with arm rests for writing and a large clock visible throughout the room.

B. Allocation of resources for different test development tasks

The different kinds of resources specified above can be used for any kind of language tests. However, the allocation of these resources differs according to the type of the test. In this proficiency test, which includes different kinds of tasks such as multiple-choice, short answer and essay writing tasks, considerable resources need to be provided not only to score the test but also to train the raters in the scoring procedures in order to ensure test usefulness.

a. Human resources

The number of human resources needed and the time they need to do their job in the administration and use stage of test development and in analyzing the feedback is specified in Table 4.1 below. It is worth mentioning here that the number of scorers, raters, administrators and other clerical support is determined by the number of test
takers. In this study, the researcher will assume the number of test takers to be one thousand, which is the average number of the test takers who take the test every year.

Table 4.1

*Human Resources in Terms of Number and Time Needed*

<table>
<thead>
<tr>
<th>Human resources</th>
<th>No. of human Resources</th>
<th>Administration</th>
<th>Scoring</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test writer and developer (the researcher)</td>
<td>1</td>
<td>3 hours</td>
<td></td>
<td>5 days</td>
</tr>
<tr>
<td>Scorers (for the reading and listening subtests)</td>
<td>4</td>
<td></td>
<td>5 days</td>
<td></td>
</tr>
<tr>
<td>Raters (for the writing subtest)</td>
<td>6</td>
<td></td>
<td></td>
<td>12 days</td>
</tr>
<tr>
<td>Raters and examiners (for the speaking subtest)</td>
<td>10</td>
<td>3 hours a day</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Administrators or proctors for the written part of the test</td>
<td>30</td>
<td>3 hours</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Support staff</td>
<td>2</td>
<td>3 hours</td>
<td>2 days</td>
<td>5 days</td>
</tr>
</tbody>
</table>

2. Space

1. For test development, the researcher should be given a suitable office.
2. For test administration and use, 15 suitable examination halls should be provided. The size of these halls should be according to the number of test takers in each hall (around 66 candidates in each hall).

C. Costs associated with the project

The costs associated with this project will be determined by the university vice rector for students’ affairs. The costs are proportional to the admission test fee collected from the test takers. This is again dependent on the number of applicants for taking the test. The test writer is normally not involved in determining the costs associated with any test administration.

D. The development time line

In high-stakes tests, preparing a time line is very important because it makes test developers aware of the needs that should be met and the time framework within which they should be completed in accordance with the objective of the test as specified by the test developer and writer (the researcher). The time line prepared by the researcher for this test is as follows:
Test Objective: to develop a test of general proficiency to be administered one month before the beginning of the four-year English programs at Taiz University. This test must be completed by the end of August 2009.

3. After getting the Ph.D. degree and going back to Yemen (January, 2009): revise the test with more experienced teachers to ensure test suitability for its purpose and for the test takers.
4. February, 2009: getting the booklet of test format and specifications ready and available for all test takers.
5. May, 2009: consult the departmental chairpersons and the deans to provide the necessary resources for fair test administration and use.

Description of TLU Domain and Task Types

The TLU domain consists of a set of settings and tasks which the students are likely to encounter in the university academic setting if they are admitted to the English programs. Before the TLU tasks are selected for use in the test, they should first be identified and described.

A. Identification of tasks

Since this test will be given to potential undergraduate students, it is appropriate at this level to focus on the language used daily in class and on campus. Extending the domain to cover the language which is used in contexts other than instructional settings would be more appropriate with upper level students. Thus, at this stage the domain is limited more or less to instructional situations and tasks.

The TLU tasks to be described are identified on the basis of the researcher’s experience of the language used in instructional settings and the professional expertise of fellow teachers.

B. Description of TLU task types

The TLU tasks are described on the basis of the framework of task characteristics presented in Table 2.2. This framework consists of a set of features for
describing five aspects of tasks: setting, rubric, input, expected response, and relationship between input and response. The purpose of this checklist is to provide a basis for language test development and use which, according to Bachman and Palmer (1996, p.47), involves three activities, which were discussed in the second chapter under the “Bachman and Palmer (1996) model”. The description of the TLU domain and task types for each of the subtests are provided below.

The TLU Domain and Task Types for the Reading Subtest

A. Characteristics of the setting

The participants are secondary school leavers who wish to major in English at Taiz University but this will depend on their achievement in this test. Other participants in the TLU domain will include teachers, who assign reading tasks and judge the outcome, and classmates and friends with whom the candidates might spend some time discussing points related to the reading task.

The physical characteristics of the TLU settings include reading at home, in classrooms, libraries and study rooms.

The time taken to read usually varies according to the type of reading. It includes daytime, evenings and weekends. Academic reading is typically slow and time consuming. For example, reading books might take more than three weeks and reading articles or short stories might take as long as three hours. In some other cases, such as reading a passage as part of a reading comprehension class, reading requires thirty minutes to one hour, and reading notices and time tables requires shorter time (not more than ten minutes).

B. Characteristics of the rubric

The characteristics of the rubric listed in the checklist are generally implicit in the TLU domain, i.e., they are not as explicit as those of the test rubric. Therefore, there is relatively little correspondence between TLU rubric and test rubric. However, there are some characteristics of the TLU rubric that can be used when describing test rubric such as the instructions. The instructions in the TLU domain can be visual (written) or aural (delivered orally) or both (visual and aural). These characteristics of the instructions in the TLU domain can also be used in the test.

C. Characteristics of the input

The TLU tasks constitute the material of the input to which language users are expected to respond. This material is described in terms of format and language characteristics.
1. Format: describes how the input is presented. It includes channel, form, language, length, type, degree of speededness, and vehicle. These components of the format of the input for the TLU reading tasks are discussed below.

   Channel: always visual
   Form: language
   Language: the target language (English)
   Length: varies depending on the academic task. In some cases, texts might be 20 pages in length or more since readers may need to read such texts in order to write an assignment or prepare for tests. In some other cases, texts might be 300 words in length or less and readers are supposed to answer some comprehension questions on them.
   Type: reading tasks for the purpose of making readers understand the reading texts and the academic materials.
   Degree of speededness: generally unspeeded

2. Language Characteristics. since the language of the input is English, the characteristics of the language of the input are related to the nature of the English language. This will be discussed in terms of language knowledge and topical knowledge.

   a. Language knowledge includes the following:

   1. Organizational characteristics which include grammatical and textual characteristics.

      • Grammatical characteristics: the language of the TLU texts is heterogeneous, sometimes simplified as in the case of introductory textbooks and sometimes complex as in the case of advanced works of literature. Vocabulary includes a wide range of general and specialized terms. The morphology and syntax include a diverse range of organized structures.

      • Textual characteristics: the language of the input is always cohesive and includes a wide range of organizational patterns.

   2. Pragmatic characteristics: according to Bachman and Palmer (1996, p. 69), pragmatic knowledge helps interpret discourse by relating statements and texts to their meanings, to the intentions of language users, and to relevant characteristics of the language use setting. Pragmatic characteristics include functional and sociolinguistic characteristics.
- Functional characteristics: the pragmatic functions are typically ideational (presenting facts) and/or manipulative (persuading the reader of a point of view). They can also be imaginative (creating fictional worlds) or heuristic (extending existing knowledge of the world).

- Sociolinguistic characteristics: the language of the input is natural and written in standard English, and the register is formal. The cultural references and figurative language vary according to the topic of the academic discipline.

b) Topical knowledge: topics in the TLU domain include academic topics such as language, literature and linguistics besides personal and cultural topics.

D. Characteristics of the expected response

Language users in any piece of discourse will typically develop a set of expectations about their respective responses. These expected responses will either be invisible (an internal interpretation of meaning), or in some form of written or spoken language. The characteristics of the expected response are also described in terms of format and language characteristics:

1. Format: includes the following components:

   Channel: may be visual such as writing a plot of a short story or oral such as making an oral summary of a chapter from a book.

   Form: language

   Language: English (the target language)

   Length: can be just a tick as in the case of choosing the correct answer in a classroom quiz or a short response such as answering gap filling or short answer questions. It can also be long as in the case of making a summary of an article or a book review.

   Type: the responses can be selected (as in multiple-choice tasks), limited (as in short answer questions) and extended (as in summarizing stories).

   Degree of Speededness: is variable; sometimes speeded but usually unspeeded depending on the purpose of the reading task

   Vehicle: live (if there is any spoken response).

2. Language Characteristics. This also includes language knowledge and topical knowledge.
a. Language Knowledge includes the following:

1. Organizational characteristics:
   - Grammatical characteristics: the grammatical characteristics of the expected response are simple and not as diverse as the language of the input.
   - Textual characteristics: these include a simple range of cohesive devices and organizational patterns.

2. Pragmatic Characteristics:
   - Functional characteristics: these are typically ideational and manipulative and in some cases imaginative depending on the demands of the input.
   - Sociolinguistic characteristics: these are the same as those of the input.

b) Topical knowledge: the characteristics of the topical knowledge are the same as those of the input.

E. Relationship between input and response

The relationship between input and response can be described in terms of reactivity, scope and the directness of the relationship.

- **Reactivity**: the relationship between input and response is generally non-reciprocal, except for the interaction between the language user and the text. This interaction is involved during the process of reading and understanding the meaning of the text.

- **Scope**: in the TLU settings, the scope of this relationship can be both broad and narrow. It is broad when the task focuses on the entire text and narrow when it focuses on specific details or parts of the text.

- **Directness**: this relationship is also both direct and indirect. It is direct in the case of understanding information explicitly stated in the text, and indirect in the case of understanding inferences and giving opinion on some events.

The TLU Domain and Task Types for the Writing Subtest

For a language test to be useful, the test tasks must correspond in demonstrable ways to language use in a specific TLU domain. The TLU tasks will be described below using the following aspects of tasks:
A. Characteristics of the setting

The participants in the TLU domain for the writing subtest are the same as those mentioned for the reading subtest.

The physical characteristics of the setting include writing at home, in classrooms and in libraries. These places may be quiet like libraries or crowded and noisy like classrooms during class discussions.

The time of writing tasks varies. It includes daytime, evenings, and weekends. Academic writing usually needs a long time. For example, writing assignments might take more than two weeks. However, some other tasks such as writing short essays might take thirty minutes, and writing memos or notes do not usually take more than ten minutes.

B. Characteristics of the rubric

The characteristics of the TLU rubric mentioned in the TLU domain for the reading subtest also pertain here.

C. Characteristics of the input

1. Format: the components of the format of the input for the TLU writing tasks are discussed below.

   Channel: can be visual or aural depending on the task. For example, in writing a summary of a classroom lecture, the input is aural and in writing a summary of a written passage, the input is visual.

   Form: is typically language. However, sometimes the form can be non-language such as pictures, tables and diagrams.

   Language: if the form is language, it will be the target language (English).

   Length: varies. It might be short such as a short prompt or very lengthy including a list of topics such as a long lecture or a short story.

   Type: tasks and prompts for the purpose of eliciting either a limited or an extended production response.

   Degree of speededness: generally unspeeded. Although there are time limits, the time is not too short to be called speeded.

   Vehicle: live (if there is any spoken input)

2. Language Characteristics: The language characteristics of the input correspond to the two components of language knowledge and topical knowledge. These components are discussed below.
a. Language knowledge

1. Organizational characteristics:
   - Grammatical characteristics: the language of the input usually contains a wide range of vocabulary and organized structures.
   - Textual characteristics: textual knowledge varies in the input. It generally includes a wide range of cohesive links and patterns of rhetorical organization.

2. Pragmatic characteristics:
   - Functional characteristics: functions also vary. They can be ideational (e.g. an opinion), manipulative (e.g. a command), heuristic (e.g. new information) and imaginative (e.g. a joke).
   - Sociolinguistic characteristics: the language of the input is usually standard, formal, and natural. Cultural references and figurative language vary according to the topic of the writing task.

b. Topical knowledge: the topical knowledge of the input varies. It can be personal, cultural, academic or technical. In TLU settings, topical knowledge can be as wide and variable as required by the English courses.

D. Characteristics of the expected response

1. Format: the format of the expected response can be described in terms of the following components.
   - Channel: visual (always written)
   - Form: language
   - Language: English (the target language)
   - Length: the responses to writing tasks in the TLU settings are usually more than two pages. In some cases such as writing memos, however, the responses are shorter.
   - Type: is extended production since language user’s response ranges from two sentences to free composition.
   - Degree of Speededness: unspeeded

2. Language Characteristics: Since the response is extended production, its language needs to be described. As mentioned earlier, the language of the response can be described in terms of language knowledge and topical knowledge.
a. Language knowledge

1. Organizational characteristics:
   - Grammatical characteristics: the vocabulary is general or specialized depending on the topic. The structural organizations follow the structures of standard English. The graphology is usually handwritten, but sometimes typewritten, as in the case of take-home assignments.
   - Textual characteristics: an acceptable range of cohesive devices and organizational patterns.

2. Pragmatic characteristics:
   - Functional characteristics: includes ideational (e.g. expressing an opinion), manipulative (e.g. presenting an argument) and possibly heuristic (e.g. giving a solution to a problem) and imaginative (e.g. using figurative language).
   - Sociolinguistic characteristics: the response is generally formal, standard and natural. It may include variable cultural references and figurative language depending on the topic of the writing task.

b. Topical knowledge: the topical characteristics of the expected response correspond to those of the input.

E. Relationship between input and response
   - Reactivity: this relationship is non-reciprocal since there is neither feedback nor interaction between language users.
   - Scope: the scope of the relationship ranges between narrow and broad depending on the input. If the input is short such as a short prompt for writing memos, the scope of the relationship will be narrow. If the input is long such as reading a short story to write its plot, the scope will be broad.
   - Directness: the relationship can be both direct and indirect. It may be direct as in the case of writing a description of a picture or writing a summary of an article, or indirect as in the case of writing on a given topic.
The TLU Domain and Task Types for the Listening Subtest

This section deals with the components of the framework of task characteristics and how they can be applied to the listening subtest.

A. Characteristics of the setting

The TLU participants include the listener and other participants like the teacher (in the classroom), classmates and friends (in classroom discussions), actors (when watching English movies as part of a literature class) and native speakers (when listening to a tape recorded conversation as part of a class on spoken English).

The physical circumstances under which listening takes place include the classroom (when listening to the lectures), the listener’s home (when listening to a tape recorded conversation as part of a take-home assignment), the campus (when listening to friends discussing some topics in English) and the lab (when watching English movies as part of a literature class). The location of listening might be quiet such as the classroom or noisy such as the campus.

The time of listening varies. It includes daytime, evenings and weekends. Listening may take a minute or an hour depending on the nature of the listening task.

B. Characteristics of the rubric

As mentioned earlier, there is little to be said about the TLU rubric. The only explicit characteristic of the TLU rubric for the listening tasks is the instructions, which are always aural and delivered in the target language (English).

c) Characteristics of the input

1. Format: includes the following:

   Channel: typically aural, but sometimes both aural (as in the case of a multiple-choice quiz which requires listening to a conversation to choose the correct written response) and visual (as in the case of English movies, which involves listening and watching the different scenes of the movies).

   Form: language

   Language: English language

   Length: the input varies. It might be short such as words, phrases or simple sentences, or long such as extended discourse.

   Type: can be a series of items (like a telephone conversation involving short utterances such as ‘yes’ or ‘really’), a task (such as a short
conversation or an extended speech like a lecture) or a prompt (like a
directive that elicits an extended response).

Degree of Speededness: usually unspeeded. However, in some cases there are
time limits as in the case of a one hour classroom lecture. In this
case, the input can be somehow speeded.

Vehicle: can be live (such as a classroom lecture) or reproduced (such as a
tape recorded conversation used in the classroom as part of a course
in spoken English).

2. Language Characteristics: this also includes the areas of language knowledge and
topical knowledge.

   a. Language knowledge

   1. Organizational characteristics:

      ■ Grammatical characteristics: the input contains a wide range of
        vocabulary; general and specialized. It also includes a variety
        of organized structures.

      ■ Textual characteristics: the input is cohesive and includes
        utterances exhibiting rhetorical organization.

   2. Pragmatic characteristics:

      ■ Functional characteristics: in the TLU domain, language
        functions vary. They can be ideational (such as descriptions,
        classifications, explanation and expression), manipulative (such
        as suggestions and warnings), heuristic (such as giving new
        information) or imaginative (such as telling jokes or reading
        poetry).

      ■ Sociolinguistic linguistic characteristics: in this TLU setting,
        the input is presented by Yemeni or Indian speakers of English.
        The register can be formal or informal depending on the
        setting. The input is usually produced naturally and is free from
        complex idiomatic expressions. Cultural references and
        figurative language vary according to the topic of the
        utterances.

   b) Topical knowledge: includes a variety of topics which may be personal,
cultural, academic or specialized.
D. Characteristics of the expected response

1. Format: includes the following components:

   Channel: can be visual (such as writing notes while listening to a lecture), or aural (such as answering the teacher’s questions orally in the classroom).

   Form: is usually language, but sometimes it can be non-language such as nodding or drawing.

   Language: is always English.

   Length: the length of the response varies considerably. It can be a tick (like in answering a multiple-choice quiz). It can be a single word or a set of simple sentences (like in responding to a classmate’s questions about a class he missed). It can also be a composition (like in writing a summary of a lecture).

   Type: the response type may be selected response (as in the case of answering a listening multiple-choice exercise in the classroom), limited production response (as in the case of taking notes on a lecture), or extended production response (as in the case of writing a composition as a summary of a lecture or a movie).

   Degree of Speededness: the response can be speeded (such as taking notes or answering questions in the classroom) or unspeeded (such as writing a summary of a lecture to an absent classmate).

   Vehicle: live, if there is any spoken response.

2. Language characteristics:

   a. Language knowledge

      1. Organizational characteristics:

         - Grammatical characteristics: the response includes a simple range of vocabulary which is usually a repetition of the vocabulary of the input. The structures of the responses, whether written or spoken, are simple.

         - Textual characteristics: the responses also include a simple range of cohesive devices and organized patterns.

      2. Pragmatic characteristics:

         - Functional characteristics: the language functions of the response can be ideational (such as expressing and explaining).
manipulative (such as suggesting and greeting), heuristic (such as using conversation to figure out an answer to a language problem posed by the teacher), or imaginative (such as making humorous comments on a classmate’s oral presentation).

- Sociolinguistic characteristics: the response is in standard English (if it is written) and in Yemeni English (if it is spoken). The register can be formal or informal depending on the input. The language of the response is natural and it may include cultural references, depending on the topic of the input.

b. Topical knowledge: The topical characteristics of the response depend directly on those of the input.

E. Relationship between input and response

- Reactivity: usually non-reciprocal, expect for example in cases like answering the teacher’s questions in the classroom or conversing with a friend. In such cases the relationship is reciprocal.
- Scope: is generally broad as in the case of listening to a long conversation for its gist. In some case where the input is short and the response is also short (as in the case of a telephone conversation), the scope of the relationship will be narrow.
- Directness: this relationship is direct since the response depends primarily on the information supplied in the input.

The TLU Domain and Task Types for the Speaking Subtest

The components of the task characteristics of speaking in the TLU domain are described below:

A. Characteristics of the setting

The participants are the speaker and his teachers, classmates and friends who are familiar to each other and can interact comfortably.

The physical characteristics include speaking in the classroom, in the teachers’ common room, in the college cafeteria and on the campus in general. These places can be quiet such as the teachers’ common room or noisy such as the campus.

The time of speaking varies. It can be minutes or hours during daytime, in the evening or on the weekend.

B. Characteristics of the rubric

The characteristics of the TLU rubric for the other skills also apply here.
C. Characteristics of the input

1. Format:
   
   Channel: usually aural (such as an oral stimulus) and sometimes visual (such as a picture which the student has to describe or a written question which has to be answered orally).
   
   Form: language and sometimes non-language such as pictures and diagrams
   
   Language: if the form is language, it is the target language (English)
   
   Length: varies. It can be single words, phrases, sentences, paragraphs or extended utterances
   
   Type: can be an item or a series of items (such as a telephone conversation) or a task (such as reading aloud as a classroom activity) or a prompt (such as role play)
   
   Degree of speededness: usually unspeeded
   
   Vehicle: live

2. Language characteristics

   a. Language knowledge

   1. Organizational knowledge:
      
      - Grammatical characteristics: general and specialized vocabulary in organized structures.
      
      - Textual characteristics: the input is cohesive and includes different rhetorical and conversational patterns.

   2. Pragmatic characteristics:
      
      - Functional characteristics: language functions of the input vary considerably including ideational (e.g. giving a description), manipulative (e.g. making a suggestion), heuristic (e.g. teaching a new language item) and possibly imaginative functions (e.g. using figurative language).
      
      - Sociolinguistic characteristics: the dialect is either Yemeni English or Indian English. The register can be formal or informal depending on the topics and the relationship between the participants (senior, equal, or junior). The language of the input is always natural with rare instances of idiomatic expressions. Cultural references and figurative language vary according to the topic.
b) Topical knowledge: covers the same range of topics as those of the TLU domains of the other skills.

D. Characteristics of the expected response

1. Format:

Channel: typically aural
Form: language
Language: English
Length: generally short including single words, phrases and sentences. However, in cases like making a presentation the response will be relatively long including extended discourse.
Type: the response can be limited production (such as a short conversation with a teacher or a friend) or extended production (such as free oral composition in the classroom or in an extracurricular activity).
Degree of Speededness: generally unspeeded. However, in classroom discussions, the response can be speeded.
Vehicle: live

2. Language characteristics:

a. Language knowledge:

   1. Organizational knowledge:

      ▪ Grammatical characteristics: include simple vocabulary which is general and sometimes specialized in a moderate range of organized structures.
      ▪ Textural characteristics: include a simple range of cohesive structures, rhetorical and conversational patterns.

   2. Pragmatic characteristics:

      ▪ Functional characteristics: include ideational functions (expressing ideas), manipulative (making apologies), and sometimes heuristic (problem-solving) and imaginative (telling a joke).
      ▪ Sociolinguistic characteristics: the dialect is Yemeni English. The register is formal when speaking to teachers or elders and informal when speaking to classmates or friends. The language of the response is generally natural without idiomatic expressions. The
use of cultural references and figurative language is less than it is in the input.

b. Topical knowledge: the response covers the same topics as the input.

E. Relationship between input and response

- Reactivity: typically reciprocal
- Scope: the scope of the relationship is usually narrow since speaking is motivated by a limited amount of input. However, in some cases (such as identifying the gist of an extended oral composition and presenting it orally), the relationship will be broad since the response is motivated by a wide amount of input.
- Directness: the relationship can be direct (such as describing the content of the picture where the information is supplied in the input) or it can be indirect (such as giving an opinion of a recent event where the speaker is expected to give new information).

Definition of the Construct(s) to be Measured

Instead of defining all the constructs and their different components together, the constructs of each of the four subtests subtest along with their specific components will be defined separately.

The Reading Subtest

Before defining the construct, it is important to note that the characteristics of the input help test developers in defining the components of the construct. Furthermore, the scores of the different test tasks should be used to measure knowledge of the different components of the construct. Therefore, these components should be specified clearly.

A. Language ability

The construct can be broadly defined as ‘the ability to read and understand textbooks, papers and other sources of information relevant to academic tasks or settings’. This definition is a theory-based definition because the test is a proficiency test that is not based no any syllabus. This definition includes the following components:

1. Ability to obtain literal information
   - ability to retrieve information explicitly stated in the text.
   - ability to identify semantic relations between words (e.g. synonymy and antonymy).
ability to understand complex and elliptical sentences, and
ability to understand cohesive links (e.g. connectives, anaphoric and cataphoric references, etc.).

2. Ability to read with interpretive comprehension
   - ability to understand implicit arguments and ideas,
   - ability to identify the purpose of the text, the setting and its target audience,
   - ability to skim the text for its central idea or ideas,
   - ability to recognize implications and presuppositions,
   - ability to make inferences and draw conclusions, and
   - ability to summarize and supply suitable titles and subtitles.

3. Ability to make personal reactions
   - ability to express opinions, and
   - ability to make predictions.

B. Strategic competence

Strategic competence, or the metacognitive strategies, according to Bachman and Palmer (p. 119), is always involved in language use and always implied in the construct definitions of language ability. Strategic competence can always be assumed to be part of the construct. However, in most language testing situations, test developers will probably not wish to make specific inferences about strategic competence. Therefore in developing this test and all its subtests, strategic competence will not be included in the construct definition.

C. Topical knowledge

Topical knowledge will not be included in the construct definition of this subtest and the other three subtests because the main interest is in making inferences about the components of language ability. Furthermore, test takers who happen to have wider and more specialized topical knowledge may be favored. To avoid this type of bias, the contributions of imagination, knowledge of the world, intelligence and any similar source of information should be kept to the minimum and topics which are relevant and common to all test takers should be preferred by test writers.

The Writing Subtest

Language ability

The writing construct can be broadly defined as ‘the ability to produce written English in response to formal academic writing tasks’. Since the purpose of this test is
to select the students who have acceptable levels of proficiency to study in English medium programs, the construct definition will be a theory-based definition. The specific components of language ability included in the construct definition are:

1. Ability to use syntactic structures
   - ability to use correct syntactic structures, and
   - ability to use appropriately formal syntactic structures.

2. Ability to use formal relevant vocabulary

3. Ability to use cohesive relationships
   - ability to make fluid and intelligible composition, and
   - ability to effectively use and vary cohesive links.

4. Ability to organize the content
   - ability to produce relevant text, and
   - ability to use correct mechanics of writing (e.g. spelling, punctuation, paragraph shape, etc.).

**The Listening Subtest**

*Language ability*

The listening construct can be defined broadly as ‘the ability to understand formal and informal stretches of spoken English in academic contexts’. Like with the definitions of the other subtests, this definition is also a theory-based definition. Because this test is a high-stakes test and because the resources available are limited, the components of the construct should be such that can be achieved and assessed reliably with the limited resources available. The components of the listening construct specified by the researcher include comprehension at different levels:

1. Comprehension at word level
   - ability to discriminate between contrastive sounds (vowels and consonants).

2. Comprehension at sentence level
   - ability to obtain explicitly stated information.
   - ability to understand elliptical structures, and
   - ability to interpret intonation patterns.

3. Comprehension at discourse level
   - ability to understand stretches of discourse (listening for the gist and for specific details).
ability to understand interactive discourse (e.g. requests, apologies, suggestions etc.), and
ability to interpret discourse markers and cohesive links.

4. Comprehension beyond discourse level
ability to infer the purpose of the speaker(s),
ability to identify the setting (the participants, the physical characteristics and the time), and
ability to infer implications and make predictions.

The Speaking Subtest

Language ability

The definition of this language ability is also a theory-based definition. The construct of speaking is broadly defined as ‘the ability to produce adequate spoken English for taking part in academic speaking contexts, both formal and informal’. This definition includes the following components:

1. Ability to use syntactic structures
   - ability to produce correct syntactic structures, and
   - ability to product appropriately formal and informal structures.

2. Ability to produce relevant and appropriate vocabulary

3. Ability to produce cohesive and relevant utterances
   - ability to use and vary cohesive markers
   - ability to produce an adequate and relevant response

4. Ability to pronounce words clearly and intelligibly.

Plan for Evaluating Qualities of Usefulness

The main purpose of this project is to develop a useful language test. Usefulness, therefore, will be an essential consideration at the different stages of test development. This can be achieved by developing a formal plan for assessing the qualities of usefulness. This plan should be considered not only in the design stage, but also during the operationalization and administration stages. This plan, according to Bachman and Palmer (pp. 133-134), is in three parts:

1. Seeking to achieve an appropriate balance among the six qualities of usefulness and setting minimum acceptable levels for each.
2. The logical evaluation of usefulness using the questions listed in the checklist for evaluating usefulness (Appendix B) to evaluate the design statement, blueprint, and test tasks.
3. Collecting qualitative and quantitative evidences during the administration stage.

The six qualities for the usefulness of this test will be discussed below using these three parts.

**Reliability**

1. Setting minimum acceptable levels of reliability

An important component in the evaluation of test usefulness is the maintaining of a balance among the qualities of usefulness and the setting of minimum acceptable levels for each of the qualities. The researcher (also the test developer) will attempt to select appropriate criteria for each level of the six qualities of test usefulness. The details of these criteria will evolve as part of the process of test development. The details of the criteria will also be refined after each administration of the test. Repeated administrations and revisions will help the researcher discover the levels for the individual qualities that can be reasonably attained given the available resources. These attainable levels are then considered for their acceptability against the overall usefulness of the test. If they are found to be unacceptable, two options remain: a) adjust the minimum acceptable levels of the qualities of usefulness, or b) reallocate the available resources.

a. Relevant considerations

1. Purpose: because this test is a high-stakes test, the minimum acceptable level for reliability will be very high. This level needs to be realistic and so two other considerations need to be discussed, namely, construct definition and the nature of the test tasks.
2. Construct definition: the construct definition focuses only on the language knowledge components of language ability and disregards the other components of strategic competence and topical knowledge. For this reason, higher levels of reliability will be expected.
3. Nature of test tasks: test tasks for each subtest are relatively uniform in their characteristics. Therefore, a high level of reliability is expected.

b. Level

A high level of reliability is expected.

c. How Specified

This level can be specified in three ways. The first way is by using appropriate reliability estimates. These estimates are a) internal consistency reliability (such as
alpha reliability) in the case of the reading and listening subtests, and b) inter-rater reliability in the case of the writing and speaking subtests. Another way of specifying this level of reliability is by the adequacy of time allocation. Sufficient time should be specified and given to finish each task. The third way of specifying this level is the clarity of scoring criteria. Scorer reliability is built into the test by having responses scored objectively in the case of the reading and listening subtests and by having responses scored using clear rating scales by two well-trained raters in the case of the writing and speaking subtests. Furthermore, to enhance the reliability of the writing subtest a third rater will compare the ratings of the two raters and make the most suitable rating.

2. Logical evaluation of reliability

A list of questions is used for the logical, or conceptual, evaluation of usefulness. Answering these questions helps the test developer develop and assess the plan for evaluating usefulness. The questions that can be asked in the logical evaluation of reliability move systematically through the task characteristics: setting, rubric, input, response and relationship between input and response. Since the essence of reliability is consistency, these questions are designed to focus on the aspects of test design that affect the consistency of test scores. These are questions 1 to 5 listed in the checklist in Appendix B. They should be answered after administering the test to provide logical evidence of reliability.

3. Procedures for collecting empirical evidence to evaluate reliability

a. Appropriate estimates of reliability

This information will be collected after test administration. The procedure for collecting information will be to calculate the internal consistency reliability (e.g. alpha reliability) for the reading and listening subtests and the inter-rater reliability for the writing and speaking subtests.

b. Adequacy of time allocation

This information will be collected during the administration stage. The procedure for collecting information will be to give a questionnaire to the test takers immediately after the test.

c. Clarity of scoring criteria

This information will be collected after test administration. The procedure for collecting this information will be a questionnaire that will be given to the test scorers and raters after scoring the test.
Construct Validity

1. Setting minimum acceptable levels of construct validity

The relevant considerations, level, and how the level is specified are hard to separate in considering construct validity.

1. Purpose: the test is a high-stakes test and therefore minimum levels of construct validity will be set quite high. That is to say, a wide range of evidence needs to be gathered in support of the validity of the score interpretations and the decisions to be made on the basis of these scores.

2. Construct definition: construct validity will be built into the test by defining the construct in a clear way that is consistent with the purpose of the test and by designing test tasks that will help in making inferences about language ability which are consistent with the construct definition. This should justify the use of the test for the intended purposes. Furthermore, evidence related to the different components of language knowledge for all the subtests needs to be collected.

3. Domain of generalization (the TLU domain): evidence supporting generalizability of score interpretations to the intended domain of generalization needs to be collected in order to show how correspondent the characteristics of the test tasks are to those of the domain of generalization. However, since the test takers are at the beginning level, it would not be appropriate at this stage to expect them to respond to tasks normally used with higher level students. Therefore, the domain of generalization will be focused on the instructional TLU domain described earlier under the description of the TLU domain and task type. Furthermore, the abilities are tested directly in order to increase the likelihood of drawing valid inferences about the construct from the test takers’ performance.

2. Logical evaluation of construct validity

The logical evaluation of construct validity can be achieved by answering questions 6 to 15 listed in the checklist of usefulness (Appendix B) after the administration of the test.
3. Procedures for collecting empirical evidence to evaluate construct validity

a) Is the construct definition adequate for the purpose of making exclusion and admission decisions?

This information will be collected before administering the test. The procedure for collecting information will be interviews with experienced teaching staff to obtain feedback on 1) the adequacy of the construct definition and the different components of language knowledge for the four subtests in relation to the purpose of the test, and 2) whether or not the scoring procedures reflect the construct definition. Information can also be collected from the entrants’ recorded performance as first year students.

b) Are the test tasks free from bias?

This information will be collected immediately after taking the test during the administration stage. The procedure for collecting information will be a questionnaire that will be given to the test takers in order to obtain feedback on the fairness of the test tasks of the four subtests with regard to the five test task characteristics (setting, rubric, input, response, and relationship between input and response).

**Authenticity**

1. Setting minimum acceptable levels of authenticity

   a. Relevant consideration

      The TLU domain is not restricted to any particular domain of language use but is rather relatively broad. Therefore, minimum acceptable level for authenticity does not have to be very high.

   b. Level

      Since the domain is relatively broad, it may be realistic to expect only a moderate level of authenticity.

   c. How Specified

      This level can be specified by doing the following:

      1. Provide a rich description of the TLU domain.

      2. Indicate the degree of correspondence between TLU tasks and test tasks. Since the language abilities tested in the four subtests are integrative in the TLU domain, the researcher provides integrative tasks in the writing and speaking subtests. The limited use of integrative tasks is due to the limited resources and the considerations of reliability.

      3. Consider the expected perceptions of test takers and test users.
2. Logical evaluation of authenticity

The logical evaluation of authenticity will be carried out after test administration by answering questions 16 and 17 listed in the checklist (Appendix B).

3. Procedures for collecting empirical evidence to evaluate authenticity
   a. To what extent do the characteristics of the test tasks correspond to those of tasks in the TLU situations?

   The stages of test development when information will be collected are the operationalization and administration stages of test development. The procedure for collecting information is comparing the description of the TLU domain and the description of the task types. Another procedure for collecting information is a questionnaire that will be given to the test takers and teachers to elicit their views on whether or not the characteristics of the test tasks correspond to those of tasks in the TLU domain.

   b. To what extent do the test developer, teachers, and students consider the authenticity of the test tasks to be ‘high’?

   This information will be collected during the operationalization and administration stages. The procedures for collecting information will be a questionnaire with a three-point (high, moderate, low) scale designed to obtain ratings of overall levels of authenticity. This questionnaire will be given to the test developer and teachers during the operationalization stage, and to test takers right after taking the test during the stage of test administration.

Interactiveness

The level of interactiveness is a function of the characteristics of the test taker that include personal characteristics, language knowledge, topical knowledge and affective schemata, and of the characteristics of the test tasks.

1. Setting minimum acceptable levels of interactiveness
   a. Relevant considerations

   The test takers’ language knowledge, topical knowledge, and affective schemata are all involved in responding to the test tasks. Of these three characteristics of the test takers, topical knowledge and affective schemata, though involved in giving a response especially in the case of the spoken response, are not included in the construct definition and therefore will not be measured or scored.
b. Level

1. Language ability: a wide range of language ability is involved. Minimum levels of interactiveness of language ability will therefore be set very high.

2. Topical knowledge: the role of the test takers’ common topical knowledge in responding to test tasks is highly involved because it is this kind of topical knowledge that is used by the test takers to respond to the input. The involvement of this type of topical knowledge can be achieved by using the type of input which presents what all test takers can identify and respond to. However, topical knowledge which is not common to all test takers will be excluded as far as possible. In either case, topical knowledge, even that which is highly involved in the test taking process, is not included in the construct definition.

3. Metacognitive strategies: it is expected that there will be involvement of strategies in accessing different bits and pieces of topical knowledge, or in deciding which of several answers to use, especially in response to items on the spoken subtest. Metacognitive strategies are, however, not included in the construct definition.

Because of the rather limited involvement of topical knowledge and metacognitive strategies, which is motivated by the purpose of the test, the construct definition, and the nature of the task, minimum levels of interactiveness will be set rather moderate.

c. How Specified

These levels can be specified by expecting the degree of involvement of the test takers’ characteristics. Language ability is expected to be highly involved, while the involvement of the topical knowledge and metacognitive strategies is expected to be limited.

2. Logical evaluation of interactiveness

For the logical evaluation of interactiveness, question 18 to 25 listed in the checklist of evaluating usefulness (Appendix B) will be answered after administering the test.

3. Procedures for collecting evidence of interactiveness

Information on interactiveness can be collected during test administration using qualitative methods such as questionnaires or interviews. These could deal with the following questions.
a. To what extent do test takers and teachers consider the involvement of language knowledge and metacognitive strategies to be high?

This information will be collected during the administration stage. The procedure for collecting information will be a questionnaire in which test takers and teachers rate the degree of involvement of language knowledge and metacognitive strategies in taking the test.

b. To what extent do test takers and teachers consider the involvement of other characteristics of test takers (topical knowledge, metacognitive strategies and affective responses) to be at least moderate?

This information will be gathered during the administration stage. For collecting information, test takers will fill in a questionnaire in which they rate their affective responses to the test tasks and the degree of involvement of topical knowledge and metacognitive strategies in taking the test.

Impact

1. Setting minimum acceptable levels of impact
a. Relevant considerations

1. Kinds of decisions to be made: high-stakes decisions. These decisions have a life-long impact on test takers and stakeholders.

2. Possible effect of using the test and of making these decisions: the test is designed to have a positive impact on test takers and instructors. The opportunity for positive impact might be increased by carefully explaining the rationale behind the test and the steps that have been taken to promote relevance, appropriateness, and fairness. However, incorrect decisions will lead to inappropriate selection and admission of students with low or insufficient level of language ability, which may have a negative impact on the students, stakeholders, programs and on society in general.

b. Level of impact

Minimum levels of impact could be set quite high because of the importance of the test both to the test takers and to the stakeholders.

c. How specified

This level can be specified by predicting the influences on the test takers, the stakeholders, the educational system, and on society. It is expected that the test takers will have positive affective responses to the test tasks because they are consistent with the instructional TLU tasks they would likely need to perform if they are admitted to
the program. The impact on the instructors is also expected to be positive because they are likely to feel that the students are being tested fairly on material that has been emphasized in the course of instruction. Naturally, such a test will also have positive impact on the educational system and on society at large.

2. Logical evaluation of impact

This evaluation can be achieved by answering questions 26 to 40 in the checklist in Appendix B after the test administration.

3. Procedures for collecting evidence of impact

a. For test takers

1. Are all test takers informed of the purpose of the test, how it will be scored, and how decisions will be made on the basis of these scores?

This information will be collected during test administration. The procedure for collecting information will be a questionnaire that will be given during test administration to all test takers in which they are asked whether they feel that they were adequately informed of the purpose, scoring and the decisions that will be made on the basis of their scores. 80 per cent of the respondents should rate this information as ‘adequate’.

2. Do 80 per cent of test takers consider the test to be ‘relatively free from bias’?

The stage of test development when this information will be collected is test administration (after taking the test). The procedure for collecting this information is the same as that for the previous question.

b) For teachers and for the educational system

1. Do teachers agree with 80 per cent of the selection decisions?

This information will be collected two weeks after students have been selected and admitted into the programs. The procedure for collecting this information will be a questionnaire that teachers fill in for each entrant.

2. Do 80 per cent of the teachers agree that the criteria for grading the test are consistent with the objectives of the courses into which the students will be selected?

This information will be collected during test use (after test administration). The procedure for collecting information will be a questionnaire that each teacher fills in at the end of each semester.

C. For Society

1. Do professionals in the community generally agree that the criteria for grading the tests are consistent with goals of society?
The stage of development when this information will be collected is the test use (after administering the test). For collecting this information, an interview will be carried out with a sample of professionals in the community (e.g. managers of private and public schools, managers of private institutes, and managers of business organizations).

**Practicality**

Practicality is an “either-or quality, with the minimum acceptable level being the threshold level at which the resources available equal or exceed the resources required” (Bachman and Palmer, p. 137).

1. Setting minimum acceptable levels of practicality

   a. Relevant considerations

      1. Threshold level of use of resources: in setting this threshold level for practicality, procedures into test development for allocating resources should be designed to insure that the resources used do not start to exceed those that have been allocated. Because the longer part of the test is scripted, relatively few resources are required to train the administrators or the proctors. The spoken subtest requires more resources for the training of the examiners. The scorers of the reading and the listening subtests do not need extensive training because of the objectivity of these two subtests. However, the raters of the writing and speaking subtests need extensive training because of the subjectivity of these two subtests. More resources are therefore required. Besides, the actual administration and scoring time will be considerable since the test will be administered on two different days, one day for the written part and the other for the spoken part. The spoken part of the test will be administered separately to each test taker. The allocation of these resources and the other types of resources is discussed above under the discussion of the “inventory of available resources and plan for their allocation and management”.

      2. Consequences of exceeding threshold level: moderate averages of resources exceeding the threshold level will not have great consequences.

2. Logical evaluation of practicality during operationalization stage

This can be carried out by answering questions 41 and 42 in the checklist of ‘The logical evaluation of usefulness’ (Appendix B). These questions can be
answered using the information on the identification and allocation of resources discussed earlier.

3. Procedures for collecting evidence of practicality

*Is the use of resources being monitored to insure that sufficient resources will be available to complete the project?*

This information will be collected before administering the test. Procedures for collecting this information include:

1. Monetary resources: depend on the number of test takers. Monetary resources will be determined by the university vice rector for students’ affairs after considering the resources needed and the number of test takers.

2. Human resources: a questionnaire to each of the teaching staff involved in test development in order to indicate the amount of human resources and the time spent on the project.

**Stage 2: Operationalization Stage**

According to Bachman and Palmer, operationalization is the process of using the components of the design statement described in the design stage to help test developers develop test tasks, a blueprint for the whole test, and finally the actual test. This section describes two interrelated activities in the operationalization stage. These two activities are:

1. developing test tasks and test task specifications, and
2. developing a blueprint for the test as a whole.

The blueprint includes the following two parts:

a. the characteristics that pertain to the structure of the test: the number of parts/tasks, the salience of parts/tasks, the sequence of parts/tasks, the relative importance of parts/tasks, and the number of tasks per part, and

b. the task specifications for each type of task that is to be included in the test.

Test task specifications include the following:

- the purpose of the test task,
- the definition of the construct to be measured,
- the characteristics of the setting of the test task,
- time allotment,
- instructions for responding to the task,
In developing the blueprint, the researcher (the test developer) determines the best way to combine and arrange the test task specifications in a test, and, at the same time, takes the qualities of usefulness into consideration by satisfying the minimum levels of these qualities that have been set under the “plan for evaluating qualities of usefulness”.

Before describing these processes of the operationalization stage, it may be pointed out that in this project, the TLU domain is not restricted and therefore it is not possible to locate a single, specific, limited TLU task type that can be used as an appropriate starting point. For this reason, the researcher develops a series of test tasks, based on the set of test task specifications that correspond in demonstrable ways to the distinctive characteristics of the TLU tasks. However, the specific characteristics of the test tasks will differ somewhat from the specific characteristics of TLU tasks. This is because certain characteristics of the TLU tasks will be selectively modified to accommodate the demands of testing.

The following is a description of the processes of the operationalization stage for the whole test, followed by a description of these processes for each of the four subtests.

**The Blueprint of the Whole Test**

**The Structure of the Whole Test**

1. Number of parts/tasks: the test is organized around two parts: a written part and a spoken part. The written part consists of three subtests (reading, writing and listening) while the spoken part, which will be tested separately, consists of the speaking subtest. The purpose of these parts is to elicit the test takers’ language ability in reading, writing, listening and speaking.

2. Salience of parts: the two parts and the four subtests within these parts are clearly distinct since they test different language abilities.

3. Sequence of parts: The order of the parts of the test is fixed. The written part will always be first and the spoken part last. The order of the subtests in the written part is also fixed. The reading subtest (a multiple-choice subtest) comes first, the writing subtest (an extended production response) comes next, and the listening comprehension subtest is administered after finishing the first two subtests. The

- characteristics of input, response, and relationship between input and response, and
- scoring method.

The following is a description of the processes of the operationalization stage for the whole test, followed by a description of these processes for each of the four subtests.
The rationale behind this order is explained by the nature of the test tasks in each subtest. The answers to the test tasks of the reading subtest are provided to the test takers, whose job is to read the passages and circle the correct answer. Such tasks may take longer time but demand less creativity on the part of the test takers. The test tasks of the writing subtest, on the other hand, require more creative work on the part of the test takers. This is why they are introduced when the test takers have completed all the tasks on the reading subtest. The tasks of the listening subtest require that all the students have finished with the first two subtests. The spoken part will be administered two weeks after the written part. This is because only the test takers who reach the adequate level of language ability (pass the written part) will take the spoken part of the test.

4. Relative importance of parts/tasks: the two parts are important and so are all the four subtests. However, the written part is more important because if test takers did not pass it, they would not be able to take the spoken part of the test.

5. Number of tasks per part: the written part consists of three subtests: reading, writing, and listening. The reading subtest consists of two tasks with fifteen items each. The writing subtest consists of two tasks, and the listening subtest consists of three tasks. The first task of the listening subtest consists of ten items, the second task consists of eight items, and the third task of four items. The spoken part, on the other hand, consists of the speaking subtest which consists of two tasks.

Test Task Specifications for the Whole Test

The purpose of the test and the construct definition: these are described clearly and in detail in the design statement under the “design stage” section.

The characteristics of the setting and those of the input and expected response: these characteristics are described in detail for each of the subtests in the description of the blueprint of the four subtests below.

The other characteristics of the whole test task specifications, which are also characteristics of the test rubric, are: time allotment, instructions and scoring method.

1. Time allotment: the time specified for the written part is three hours and the time for the spoken part is twelve minutes for each test taker. The full time for the spoken part will depend on the number of the test takers who pass the written part. However, this time should not exceed three hours a day so that the examiners do not feel bored or tired and they retain their concentration and interest throughout.
2. **Instructions:** The instructions for the whole test as well as for its subtests will be provided in the booklet that will be made available for the candidates five months before taking the test. This booklet will help them prepare for the test and answer their questions about the whole test. The instructions that will be provided in the booklet include descriptions of the following: test purpose, language abilities to be tested, parts of the test and their relative importance, procedures to be followed for all parts of the test, and the scoring method.

These instructions are important because they make clear to the test takers everything they would need to know about the test: the exact nature of the testing procedures and of the test tasks, how they are to respond to these tasks, and how their responses will be evaluated. These instructions will be written in clear and simple English in order to make them easy to understand by test takers with different language abilities. Above all, effective instructions help assure test takers that the test is relevant, appropriate and fair.

3. **Scoring method:**
   
   a. **Criteria for correctness:** The test consists of tasks that require a right/wrong method of scoring in the case of the reading and listening subtests, and criterion-referenced language ability scales in the case of the writing and speaking subtests. The total score of the whole test (100) will be formed by aggregating the total scores of the reading and listening subtests, and the composite scores of the writing and the speaking subtests. The pass mark or the cut-off score for each subtest is 16 out of 30 for the reading subtest, 12 out of 24 for the writing subtest, 12 out of 22 for the listening subtest, and 12 out of 24 for the speaking subtest. Accordingly the cut-off score of the written part of the test is (40 out of 76). This cut-off score is set at a clearly established level of ability obtained from the cut-off scores of the reading, writing, and listening subtest. The test takers who pass the written part, i.e. those who score 40 and above, will be given the spoken part of the test. If they pass it, i.e. if they get past the cut-off score set at 12 marks, they will reach the cut-off score of the whole test (52 marks) and consequently they will be admitted to the programs.

   b. **Procedures for scoring the response:** The objective subtests, reading and listening, will be scored by one or two scorers each, depending on the number of test takers. The subjective subtests, writing and speaking, will be rated by
two raters each. For the writing subtest, a third rater will determine the final rating of the subtest after comparing the ratings of the two raters. The total scores of the subtests will be aggregated. These total scores of each subtest will also be aggregated to form the total score of the whole test (100 marks). These aggregations will be carried out by a different scorer whose job is to aggregate item, task, subtest and test part scores into the total test score.

c. **Explicitness of criteria and procedures:** Test takers will be informed clearly about the nature of the scoring criteria and the procedures for scoring the responses in the booklet of test specifications that will be given to them before taking the test.

**The Blueprint of the Four Subtests**

**The Blueprint of the Reading Subtest**

**The structure of the reading subtest.** The structure of the reading subtest is described in terms of the following:

1. **Number of parts/tasks:** the reading subtest consists of two reading passages (321 and 494 words long respectively) and fifteen multiple-choice items for each passage.

2. **Salience of parts:** the two passages are clearly distinct. One of them is an extract from a non-fiction book while the other is an extract from an English novel. They deal with two different topics. The purpose of using two passages is to help provide more realistic and reliable means of assessment.

3. **Sequence of parts:** since the two passages are distinct, any order should be possible. However, the shorter passage, the extract from the non-fiction book, will be presented first in order to make test takers feel at ease.

4. **Relative importance of parts or tasks:** the two passages carry the same importance. Each of them can stand alone and each of them tests reading ability at the literal and referential or interpretive levels of comprehension in two different topics.

5. **Number of tasks per part:** two reading passages. Each reading passage is followed by fifteen multiple-choice items with three choices for each item.

**Test task specifications for the reading subtest.** The test task specifications for the reading subtest are described in terms of the following:

1. **Purpose:** as mentioned in the design statement, this test in its two parts and four subtests will be used to achieve a single purpose. This purpose is to select the most qualified students for admission into the English programs offered by the English departments at Taiz University. The four subtests are, therefore, used to achieve the
same purpose. This subtest will contribute to making the best selection with regards to the reading ability. The purpose of the reading subtest is to test the test takers’ ability to understand written English and answer related comprehension questions.

2. Definition of the construct: the two tasks of the reading subtest are developed from the construct definition. The fifteen multiple-choice questions on each passage test the test takers’ ability to obtain literal information, to read with interpretive comprehension and to make individual reactions to the ideas in the passages. These are the reading abilities included in the construct definition of this subtest.

3. Setting: the characteristics of the setting are described in Table 4.2 below.

4. Time allotment: the total time needed to respond to this subtest is one and a half hours (forty five minutes for each task).

5. Instructions: as mentioned earlier, the full description of the test instructions will be provided to all test takers in the booklet of test specification. The instructions for this subtest, however, are written in simple English. They are intended to be short, clear and sufficiently detailed for test takers to know exactly what they are expected to do. Specific instructions are written for each passage and the score of the items and the scores of the whole task as well as the cut-off score of each task are written at the end of each task. These instructions are provided in the test itself (Appendix I).

6. Characteristics of input and expected response: these characteristics are described in Table 4.2 below.

7. Scoring method:

   a. Criteria for correctness: The two tasks of the reading subtest are selected response tasks. Therefore, they will be scored according to a single criterion. The method of scoring for this task type is a right/wrong scoring method. The maximum score for this subtest is thirty (fifteen marks for each task) which means that one mark will be assigned to each multiple-choice item. The cut-off score is sixteen (eight mark for each task). The items on each task are arranged thus: four easy, seven moderate and four difficult items. The cut-off score is specified by aggregating the scores of all the easy items and four of the moderate items.

   b. Procedures for scoring the response: all answer scripts will be scored by one or two scorers. Since the scoring of this subtest is a right/wrong scoring method, there will be no variations in the scoring. The scores of each test taker in the two tasks will be aggregated by a different scorer to form the total score of the reading subtest.
**C. Explicitness of criteria and procedures:** the test takers will be informed about the scoring criteria and the procedures for scoring the responses in the booklet of test specifications before taking the test.

The following table lists the test task characteristics for the reading subtest.

**Table 4.2**

**Task Characteristics for the Reading Subtest**

<table>
<thead>
<tr>
<th></th>
<th>TEST TASK 1</th>
<th>TEST TASK 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SETTING</strong></td>
<td>American Sign Language</td>
<td>Tender Troubles</td>
</tr>
<tr>
<td><strong>Participants</strong></td>
<td>Test takers; paid test administrators in the university testing center, experienced and trained in test administration.</td>
<td>Same as for Task 1</td>
</tr>
<tr>
<td><strong>Time of task</strong></td>
<td>8.00 a.m. August 15\textsuperscript{th}</td>
<td>8.45 a.m. August 15\textsuperscript{th}</td>
</tr>
<tr>
<td><strong>TEST RUBRIC</strong></td>
<td>Discussed above under the structure and task specifications of the reading subtest.</td>
<td>Same as for Task 1</td>
</tr>
<tr>
<td><strong>INPUT</strong></td>
<td><strong>Format</strong></td>
<td><strong>Channel</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Language</strong></td>
<td>English (target language)</td>
</tr>
<tr>
<td></td>
<td><strong>Length</strong></td>
<td>A reading passage which consists of 321 words</td>
</tr>
<tr>
<td></td>
<td><strong>Type</strong></td>
<td>A task that requires understanding the written passage and answering some comprehension questions</td>
</tr>
<tr>
<td></td>
<td><strong>Spedeedness</strong></td>
<td>Speeded</td>
</tr>
</tbody>
</table>

152
**Textual**
- Syntax: standard English.
- Graphology: typewritten.
- Cohesion: wide range of cohesive devices. Rhetorical organization: wide range of organizational patterns dealing with research on American Sign Language by the American linguist Stokoe.

**Pragmatic characteristics**
- Ideational and heuristic

**Sociolinguistic**
- Cultural reference and figurative language: none.

**Topical knowledge**
- Sign Language

**EXPECTED RESPONSE**

**Format**
- Channel: Visual
- Form: Non-language
- Language: N/A
- Length: N/A
- Type: Selected response
- Speededness: Speeded

**Language characteristics**
- Since the response is a selected response, there is no language to be described.

**RELATIONSHIP BETWEEN INPUT AND RESPONSE**

**Reactivity**
- Non-reciprocal

**Scope of relationship**
- Broad and narrow: broad when test takers are required to read the whole passage in order to, for example, choose the main idea or determine inferences; narrow when they are required to scan the passage for specific details such as finding the referent of a pronoun or finding a synonym for a word or phrase from a specified line or paragraph.
- Direct

**Directness of relationship**
- Direct
The Blueprint of the Writing Subtest

The structure of the writing subtest. The structure of the writing subtest is described in terms of the following:

1. Number of parts/tasks: this subtest consists of two tasks – an integrated task and an independent task. The integrated task requires test takers to rewrite a written report of an incident. The independent task requires them to write about a topic. The purpose of including an integrated task in the test is that in the TLU domain there is integration between writing and reading. Many writing tasks are based on written texts.

2. Salience of parts: the two tasks are clearly distinct. They deal with different topics.

3. Sequence of parts: the integrated task is introduced first. The task requires the test takers to re-write a passage they have in front of them and so is less demanding. The independent task, on the other hand, makes more demands on the test takers because it requires them to create a written passage on the basis of a prompt. It is introduced after the integrated task.

4. Relative importance of parts or tasks: the two tasks are important. Each one of them can stand alone. However, the independent task is given more time and more marks since the test takers are required to create a passage while the integrated task require the test takers to rewrite a passage which is already given to them.

5. Number of tasks per part: two tasks.

Test task specifications for the writing subtest. The test task specifications for the writing subtest are described in terms of the following:

1. Purpose: the purpose of this subtest is to select for admission into the English programs the test takers who have acceptable levels of ability to produce adequate written English for academic writing tasks.

2. Definition of construct: the two tasks of this subtest are based on the construct definition of the writing subtest and so are the rating scales used in the rating of test takers’ responses to these two tasks.

3. Setting: the characteristics of the setting for this subtest are shown in Table 4.3.

4. Time allotment: twenty minutes for the integrated task and thirty minutes for the independent task.

5. Instructions: as mentioned earlier, the full description of the instructions will be provided in the booklet of test specifications. However, for ease of reference, a brief description of the instructions for the writing subtest is given here. These instructions are written briefly in simple English. They are made clear and sufficient to ensure that
all test takers understand what is expected from them. There are specific instructions for each task, along with the total score and the cut-off score of that task. These instructions are also written in the test itself (Appendix I).

6. Characteristics of input and expected response: these characteristics are described in Table 4.3 below.

7. Scoring method:

   a. Criteria for correctness: the criteria for correctness are criterion-referenced language ability scales based upon the theory-based definition of the ability to use language in writing tasks. The researcher uses the rating scales described in Bachman and Palmer but changes the components of the scale to suit this particular group of test takers and this particular testing situation. The test takers’ writings will be scored from 1 to 4 on the ability to use syntactic structures, the ability to use formal and relevant vocabulary, the ability to use cohesive relationships, and the ability to organize the content. According to Bachman and Palmer (1996, p. 269), the zero level is included in each scale but none of the test takers are at this level. The actual scales to be used are provided in Appendix G. The maximum score for the first task is eight marks and for the second task is sixteen marks (twenty four for this subtest) and the cut-off score is four marks for the first task and eight marks for the second task (twelve marks for this subtest). The cut-off score is specified by aggregating the marks allocated to the moderate level of ability on the four rating scales (2 marks). For the first task, the cut-off score is $2 \times 4 = 8 / 2 = 4$ marks. The score is divided by two because the question is assigned eight marks. The cut-off score of the second task is $2 \times 4 = 8$ marks.

   b. Procedures for scoring the response: all compositions are read and rated independently by two trained raters. The raters review the scales before each rating session. They assign ratings on the four scales and then aggregate the scores from different analytic rating scales to form the composite score. For the first task, the composite score will be formed by adding up the scores from the different analytical rating scales (16 marks) and dividing the total by two (8 marks). For the second task, the composite score will be formed by adding up the scores from the different analytical rating scales (16 marks). After the rating of all the compositions by the two raters, these compositions will be rated by a third rater who will compare the two ratings of the two raters until the third rater arrives at
the best rating of the compositions. The use of a third rater enhances the reliability of the scores and the fairness of the decisions based on these scores.

C. Explicitness of criteria and procedures: The criteria for correctness and the procedures for scoring the responses will be made clear to all the test takers in the booklet of test specifications, which means that the criteria and procedures of scoring will be explicit for all test takers before they take the test.

The table of test task characteristics for the writing subtest is shown below.

Table 4.3
Task Characteristics for the Writing Subtest

<table>
<thead>
<tr>
<th>.setting</th>
<th>TEST TASK 1 (integrated task)</th>
<th>TEST TASK 2 (independent task)</th>
</tr>
</thead>
<tbody>
<tr>
<td>setting</td>
<td>Re-writing a report</td>
<td>Writing an essay</td>
</tr>
<tr>
<td>Participants</td>
<td>Test takers; paid test administrators in the university testing center, experienced and trained in test administration.</td>
<td>Same as for Task 1</td>
</tr>
<tr>
<td>Time of task</td>
<td>9.30 a.m. August 15th</td>
<td>9.50 a.m. August 15th</td>
</tr>
<tr>
<td>TEST RUBRIC</td>
<td>Discussed above under the structure and task specifications of the writing subtest.</td>
<td>Same as for Task 1</td>
</tr>
<tr>
<td>INPUT</td>
<td>Format</td>
<td>Visual (written)</td>
</tr>
<tr>
<td>Form</td>
<td>Language</td>
<td>Language</td>
</tr>
<tr>
<td>Language</td>
<td>English (the target language)</td>
<td>English (the target language)</td>
</tr>
<tr>
<td>Length</td>
<td>A written passage which consists of 200 words. Input used in writing the answer: medium (one page).</td>
<td>Prompt short. Input used in writing the answer: medium (one page).</td>
</tr>
<tr>
<td>Type</td>
<td>Prompt and task</td>
<td>Prompt</td>
</tr>
<tr>
<td>Speededness</td>
<td>Speeded</td>
<td>Speeded</td>
</tr>
<tr>
<td>Language characteristics</td>
<td>156</td>
<td>156</td>
</tr>
</tbody>
</table>
## Organizational characteristics

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Textual</td>
<td>Cohesion: cohesive text. Rhetorical organization: a text exhibiting a range of organized patterns dealing with an incident which relates to a young man saving a drowning young boy.</td>
<td>Cohesion: cohesive prompt. Rhetorical organization: a few procedures for writing the essay.</td>
</tr>
</tbody>
</table>

## Pragmatic characteristics

<table>
<thead>
<tr>
<th>Functional</th>
<th>Ideational and manipulative</th>
<th>Ideational and manipulative</th>
</tr>
</thead>
</table>

## Topical knowledge

| Water incident | Making university education in Yemen free of charge |

## EXPECTED RESPONSE

<table>
<thead>
<tr>
<th>Format</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Channel</strong></td>
</tr>
<tr>
<td><strong>Language</strong></td>
</tr>
<tr>
<td><strong>Type</strong></td>
</tr>
<tr>
<td><strong>Speededness</strong></td>
</tr>
</tbody>
</table>

## Language characteristics

<table>
<thead>
<tr>
<th>Organizational characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Textual</strong></td>
</tr>
</tbody>
</table>

## Functional

<table>
<thead>
<tr>
<th>Ideational and possibly</th>
<th>Ideational and possibly</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topical knowledge</td>
<td>Water incident. Making university education in Yemen free of charge.</td>
</tr>
</tbody>
</table>

### RELATIONSHIP BETWEEN INPUT AND RESPONSE

| Reactivity | Non-reciprocal | Non-reciprocal |
| Scope of relationship | Relatively broad (there is much input for the test takers to process in order to give their responses). | Relatively narrow (there is a limited amount of input in the prompt). |
| Directness of relationship | Direct (the response includes primarily information supplied in the input). | Indirect (the response includes information not supplied in the input). |

### The Blueprint of the Listening Subtest

**The structure of the listening subtest.** The structure of the listening subtest is described in terms of the following:

1. **Number of parts/tasks:** the subtest is organized around three tasks: comprehension of statements, comprehension of mini-dialogues and comprehension of mini-talks. The purpose of using these three types of tasks is to get test takers to demonstrate their ability to comprehend utterances of different lengths.

2. **Salience of tasks:** these tasks are clearly distinct, since each of them tests the test taker’s ability at a different level.

3. **Sequence of tasks:** the order of these tasks moves from the shortest to the longest. Statement comprehension comes first, then mini-dialogue comprehension, and finally mini-talks comprehension. The introduction of the shorter utterance first is meant to help the test takers get used to the accent and voice quality of the speaker or speakers before the longer utterances are introduced.

4. **Relative importance of tasks:** all the tasks carry the same importance because in this test, as with TLU settings, these tasks can be used alone.

5. **Number of tasks per part:** three tasks. The first task consists of ten statements, the second task consists of eight mini-dialogues and the last task consists of two mini-talks and two items for each mini-talk.
Test task specifications for the listening subtest. The test task specifications for the listening subtest are described in terms of the following:

1. Purpose: the purpose of the listening subtest is to select for admission into the English programs the test takers who demonstrate ability to comprehend spoken dialogues, discussions and lectures delivered in English.

2. Definition of construct: the construct and its components are defined in the design statement. The tasks in this subtest are designed on the basis of the construct definition of this subtest.

3. Setting: the characteristics of the setting are described in Table 4.4 below.

4. Time allotment: the time specified for the subtest is forty minutes: ten minutes for task 1, fifteen minutes for task 2, and fifteen minutes for task 3.

5. Instructions: Besides the clear explanation of the test instructions provided in the booklet of test specifications, the instructions for this subtest have the following characteristics:
   
   a. Language: simple and clear English which is familiar to all test takers.
   
   b. Channel: the instructions will be presented orally.
   
   c. Instructions: these instructions are included in the test itself (Appendix I).

6. Characteristics of input and expected response: these characteristics are described in Table 4.4 below.

6. Scoring method:

   a. Criteria for correctness: the criteria for scoring this subtest are the same as those for the reading subtest since the tasks in both these subtests are selected response tasks which require a right/wrong scoring method. The maximum score of the listening subtest is twenty two: ten for task 1, eight for task 2, and four for task 3. One mark is given to each item in each task. The cut-off score for this subtest is twelve (six for task 1, four for task 2, and two for task 3). The items on task 1 are arranged thus: three easy, five moderate and two difficult items. The cut-off score is specified by aggregating the scores of all the easy items and three of the moderate items. The items on task 2 are arranged thus: two easy, four moderate and two difficult items. The cut-off score is specified by aggregating the scores of all the easy items and two of the moderate items. The cut-off score of the third task is specified by aggregating the scores of the easier two of the four items.
b. Procedures for scoring the response: the three tasks will be scored by one scorer (they may also be scored by two scorers if needed without affecting the consistency of the scoring). The scores of each test taker on all the three tasks will then be aggregated by another scorer to form the total score of this subtest.

c. Explicitness of criteria and procedures: test takers will be informed about the criteria and the procedures of scoring their responses before taking the test. This information will be provided in the booklet of test specifications.

The table of test task characteristics for this subtest is given below:

Table 4.4
Task Characteristics for the Listening Subtest

<table>
<thead>
<tr>
<th></th>
<th>TEST TASK 1</th>
<th>TEST TASK 2</th>
<th>TEST TASK 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SETTING</strong></td>
<td>Statements</td>
<td>Mini-dialogues</td>
<td>Mini-talks</td>
</tr>
<tr>
<td><strong>Physical Characteristics</strong></td>
<td>Location: university examination halls with speakers fitted at the front and back sides of the halls. Noise level: low. Temperature and humidity: comfortable. Seating conditions: individual chairs with arm rests for writing. Lighting: well-lit. Material and equipment: tape recorders, booklets, pens, pencils with erasers and pencil sharpeners.</td>
<td>Same as for Task 1</td>
<td>Same as for Task 1</td>
</tr>
<tr>
<td><strong>Participants</strong></td>
<td>Test takers; paid test administrators in the university testing center, experienced and trained in test administration.</td>
<td>Same as for Task 1</td>
<td>Same as for Task 1</td>
</tr>
<tr>
<td><strong>Time of task</strong></td>
<td>10.20 a.m. August 15th</td>
<td>10.30 a.m. August 15th</td>
<td>10.45 a.m. August 15th</td>
</tr>
<tr>
<td><strong>TEST RUBRIC</strong></td>
<td>Discussed above under the structure and task specifications of the listening subtest.</td>
<td>Same as for Task 1</td>
<td>Same as for Task 1</td>
</tr>
<tr>
<td><strong>INPUT</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Format</strong></td>
<td><strong>Channel</strong></td>
<td><strong>Form</strong></td>
<td><strong>Language</strong></td>
</tr>
<tr>
<td>------------</td>
<td>-------------</td>
<td>----------</td>
<td>--------------</td>
</tr>
<tr>
<td></td>
<td>Aural (spoken) and visual (written).</td>
<td>Aural (spoken) and visual (written).</td>
<td>Aural (spoken) and visual (written).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Language</td>
<td>Language</td>
</tr>
</tbody>
</table>
The Blueprint of the Speaking Subtest

The structure of the speaking subtest. The structure of the speaking subtest is described in terms of the following:

1. Number of parts/tasks: this subtest is organized around two tasks: oral interview between the examiner and the test taker, and a task that require the test taker to describe the content of a picture. The purpose of the oral interview task is to elicit the test taker’s ability to communicate with people in the academic settings while the purpose of the description task is to measure the test takers’ ability to describe things they see in their surroundings (as exemplified in the pictures).

2. Salience of tasks: the two tasks are clearly distinct. They do not depend on each other.
3. **Sequence of tasks**: the oral interview will be carried out first because it involves ‘warming up’ the test takers in order to make them feel relaxed and comfortable before they set out to describe the pictures.

4. **Relative importance of parts/tasks**: the two tasks are equally important because communicating and describing are important activities in any academic setting.

5. **Number of tasks per part**: two tasks. The first task consists of some questions and responses which are necessary to help the examiner locate the test taker in the scale that best describes his/her speaking ability. The second task consists of one picture which the test taker has to describe.

**Test task specifications for the speaking subtest.** The test task specifications for the speaking subtest are described in terms of the following:

1. **Purpose**: since the purpose of the whole test is to select the qualified test takers, this subtest will contribute in making the best selection of the test takers with regards to their speaking ability. Only those test takers who reach an acceptable level of speaking ability that enables them to communicate with others and to describe will be selected and admitted into the offered courses.

2. **Definition of construct**: this definition is given in the design statement. The two tasks of this subtest are designed to measure the different components of this construct and the scales used for rating these tasks are also based on the definition of the construct and its components.

3. **Setting**: the characteristics of the setting are described in Table 4.5 below.

4. **Time allotment**: each test taker is given ten minutes for the first task and two minutes for the second task.

5. **Instruction**:
   a. **Language**: simple and clear English
   b. **Channel**: delivered orally by the examiner
   c. **Instructions**: the full description of the instructions for this subtest, along with the other subtests, will be provided to the students before taking the test. However, for this project, the instructions for this subtest are also provided in the test itself (Appendix I).

6. **Characteristics of input and expected response**: the characteristics of input and expected response are described in Table 4.5 below.
7. Scoring method:

a. **Criteria for correctness:** the criteria for correctness are criterion-referenced, language ability scales based upon theory-based definition of the ability to use language in speaking tasks. The test takers’ responses will be scored from 1 to 4 on their ability use syntactic structures, to produce relevant and appropriate vocabulary, to produce cohesive and relevant utterances, and to pronounce words clearly and intelligibly. The actual scales to be used are provided in Appendix H. The maximum score for this subtest is twenty four marks (sixteen for the first task and eight for the second task) and the cut-off score is twelve marks (eight for the first task and four for the second task). The cut-off score is specified by aggregating the marks allocated to the moderate level of ability on the four rating scales (2 marks). The cut-off score of the first task is $2 \times 4 = 8$ marks. For the second task, the cut-off score is $2 \times 4 = 8 / 2 = 4$ marks. The score is divided by two because the question is assigned eight marks.

b. **Procedures for scoring the response:** there will be two examiners: the primary examiner who conducts the test, and the secondary examiner who acts as an observer and assists with the rating. These two examiners rate the two tasks separately, both during and immediately following the test. If the examiners disagree by more than one level on each scale, they discuss their notes in order to arrive at a rating that they can agree upon. After assigning the ratings of the two tasks on the four scales, the scores from the different analytical rating scales will be aggregated to form the composite scores of the tasks. For the first task, the composite score will be formed by adding up the scores from the different analytical rating scales (16 marks). For the second task, the composite score will be formed by adding up the scores from the different analytical rating scales (16 marks) and dividing the total by two (8 marks).

c. **Explicitness of criteria and procedures:** The test takers will be informed clearly and in detail about the scoring criteria and the procedures for scoring the responses in the booklet of test specifications.

The following is a table of the test task characteristics for the speaking subtest.
Table 4.5

Task Characteristics for the Speaking Subtest

<table>
<thead>
<tr>
<th>SETTING</th>
<th>TEST TASK 1</th>
<th>TEST TASK 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Physical Characteristics</strong></td>
<td>Oral interview</td>
<td>Pictorial description</td>
</tr>
</tbody>
</table>

| Time of task | 9.00 a.m. August 29th (for the first test taker on the first day). | 9.10 a.m. August 29th (for the first test taker on the first day). |

| TEST RUBRIC | Discussed above under the structure and task specifications of the speaking subtest. | Same as for Task 1 |

<table>
<thead>
<tr>
<th>INPUT</th>
<th>TEST TASK 1</th>
<th>TEST TASK 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Format</strong></td>
<td>Aural (the examiner’s conversational turns).</td>
<td>Both aural (the examiner’s questions about the content of the picture) and visual (the picture to be described).</td>
</tr>
<tr>
<td><strong>Channel</strong></td>
<td>Language (the target language)</td>
<td>Both language and non-language (the picture)</td>
</tr>
<tr>
<td><strong>Form</strong></td>
<td>English (the target language)</td>
<td>English (the target language)</td>
</tr>
<tr>
<td><strong>Language</strong></td>
<td>Short to medium (depending on the interactiveness of the test takers).</td>
<td>Short (simple questions about the pictures).</td>
</tr>
<tr>
<td><strong>Length</strong></td>
<td>Prompt: examiner asks questions and/or gives responses.</td>
<td>Prompt: the picture and the examiner’s questions on the picture.</td>
</tr>
<tr>
<td><strong>Type</strong></td>
<td>Live</td>
<td>Highly Speeded</td>
</tr>
<tr>
<td><strong>Speededness</strong></td>
<td>Highly Speeded</td>
<td>Live</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Language characteristics</th>
<th>TEST TASK 1</th>
<th>TEST TASK 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Organizational characteristics</strong></td>
<td>Vocabulary: moderate range of general vocabulary related to the test taker’s life, interests and</td>
<td>Vocabulary: very narrow range of vocabulary related to the pictures.</td>
</tr>
<tr>
<td><strong>Grammatical</strong></td>
<td>Morphology and syntax: simple.</td>
<td></td>
</tr>
</tbody>
</table>

**Textual**  
Cohesion: textually cohesive.  
Rhetorical organization: individual statements requiring test takers to talk about their lives and the topic of their choice; the examiner's interactive turns in the conversation. 

**Pragmatic characteristics**  
**Functional**  
Ideational and manipulative  
Dialect/variety: standard. Register: both formal and informal.  
Naturalness: natural. Cultural reference and figurative language: vary according to the themes of the questions. 

**Sociolinguistic**  

**Topical knowledge**  
Personal and academic topics. General in the form of a picture) 

**EXPECTED RESPONSE**  

**Format**  
Channel: Aural  
Form: Language  
Language: English (the target language)  
Length: Relatively short: individual to slightly extended utterances.  
Type: Limited and slightly extended production.  
Speededness: Highly speeded  
Vehicle: Live  

**Language characteristics**  
**Organizational characteristics**  
Grammatical  
Vocabulary: general vocabulary. related to the content of the picture. Morphology and syntax: simple organization.  

**Textual**  
Cohesion: cohesive throughout. Rhetorical organization: limited to extended description. 

**Pragmatic characteristics**  
**Functional**  
Ideational and manipulative  
Dialect/variety: standard. Register:  

**Sociolinguistic**  
Dialect/variety: standard. Register:
At the end of this stage, the preliminary version of the test will be ready for try-out. However, trying-out or pre-testing the test at Taiz University is not possible because of the limited amount of resources and because trying-out the test would risk making the content of the test known to those who are to take it. Therefore, this version will not be tried-out. It will rather be the actual test that will be administered. This version of the test is provided in Appendix I.

**Stage 3: Test Administration and Use**

**Test Administration**

The first two sections of this chapter described the first two stages of test development namely, test design and operationalization. In these two stages, the researcher developed the design statement, the blueprint, and the preliminary version of the test. This section focuses on the third stage of test development, i.e. test administration. Administration typically takes place in two phases: try-out and operational testing. However, in this project, trying-out is not possible for the reasons mentioned above. According to Bachman and Palmer, the administration stage involves giving the test to a group of test takers, collecting information and analyzing this information. Analysis of this information is important for two purposes: 1) assessing the usefulness of the test, and 2) making the inferences or decisions for
which the test is intended. To achieve the first purpose, feedback from a variety of sources will be collected to enable the researcher to confirm or revise the original design statement, the test blueprint and test task specifications as needed, and carry out the assessment of usefulness. The second purpose will be achieved by relying primarily on the scores obtained from the test itself.

The following section discusses the procedures for administering the test and collecting feedback for assessing usefulness.

**Procedures for Administering the Test**

To accomplish the two purposes of test administration mentioned above, it is necessary to have some control over the procedures for administering the test. These procedures include the following:

A. **Preparing the testing environment**

The testing environment will be prepared by the personnel at the university testing centre in accordance with the procedures specified in the blueprint. The test developer will go over the instructions with the test administrators at the testing center to make sure they understand them and answer any question they may have. This should be done to maximize all the qualities of usefulness of the test.

B. **Communicating the instructions**

After giving the test to the test takers, the second step is to make all test takers understand the instructions. This will be achieved by having the test administrators read the instructions aloud while the test takers look at them in their test booklets.

C. **Maintaining a supportive testing environment**

Maintaining a supportive testing environment throughout the test includes avoiding distractions such as those that take place due to noise and excessive movement. This would be maintained by requiring test takers to remain in their seats until the end of the time specified for the test to avoid movement and noise. A supportive testing environment will also be maintained by making sure that the testing center keeps on hand well-trained proctors, who have enough information on the test, to answer the test takers’ questions about the instructions or the testing procedures.

D. **Collecting the test**

Test administrators will collect all testing materials. They will not hurry to collect the tests, but they will rather give the test takers time to fill in the questionnaire given to them. While the test takers are providing this feedback, the
proctors can quietly collect the test booklets. After finishing with the questionnaires, test takers can then leave at their own pace.

Collecting Feedback for Assessing Usefulness

As indicated above, the primary purpose of collecting feedback is to provide information relevant to evaluating the qualities of usefulness and to making revisions to the test. This section includes a discussion of the specific purposes for which feedback can be collected, the kinds of feedback that can be collected, the different sources of this feedback, the resources involved in obtaining feedback, and the methods that will be used for obtaining feedback.

1. Specific uses of information collected for assessing test usefulness

The source of information for considering test usefulness will change according to the stage of test development. During the stages of test design and test operationalization, the primary source of information would be the test developer’s (the researcher) informed judgments. For the third stage, the stage of test administration, the test developer gets to obtain information from additional sources other than her own judgments. These additional sources of information make possible an initial evaluation of the usefulness of the test, based on empirical data and observations, and allow modifications in the test or the testing procedures to improve its usefulness. There are a number of specific uses of the information from test administration. The uses of this information are described as follows.

a. Getting information about administrative procedures: this can be achieved by giving a questionnaire to test administrators and test takers to collect information about problems with administrative procedures, and opinions on whether these are related to the testing environment or to the procedures for giving the test. These problems will be solved and avoided in subsequent tests.

b. Determining appropriate time allocations: this will be achieved by giving a questionnaire to the test takers to know their opinions about time allocations and about whether allocations of time allow test takers to perform at their best.

c. Identify problems in task specifications and clarity of instructions: this will also be achieved using a questionnaire that asks test takers about their opinions on task specifications and instructions.

d. Obtaining preliminary information on how test takers react to the test: this is the last use of the information obtained from test administration. This

169
information relates to how test takers respond to the test tasks in three areas: test-taking processes, perceptions of test tasks, and test performance.

Information on test-taking processes and on perception of test tasks will be obtained using qualitative assessment procedures. Feedback on test-taking processes will be obtained by observing test takers as they take the test and by using a questionnaire that will be given to the test takers to obtain information on their experience on test taking processes. The test takers’ perception of test tasks will also be obtained from a questionnaire that asks test takers about their perception of the test tasks and their difficulty, the appropriateness of the time allocation and administration procedures, and the fairness of the test as a whole. Information on the test takers’ performance, on the other hand, will be obtained from the test takers’ scores. The test taker’s scores on individual items, tasks, subtests and the test as a whole will be collected and analyzed statistically in order to find out the difficulty level of the individual items, tasks, subtests and the whole test, the degree to which individual tasks function in relation to other tasks, the reliability of the test, and so forth. These scores will be used to help make judgments about whether the test is providing the information it was designed to provide.

2. Kinds of feedback

The feedback sought is of two kinds:

a. Feedback about test takers’ language ability: feedback about the test takers’ language ability includes information on the extent to which the test tasks require the test takers to use their organizational and pragmatic knowledge (the two components of language ability) and their topical knowledge. This kind of information is useful in making a preliminary assessment of the construct validity, authenticity, and interactivity of the test tasks. Besides, additional feedback information will be gathered from qualitative feedback (obtained from observation of the test takers’ responses to the test tasks, from the analysis of the test takers’ responses, or from the questionnaire filled by the test takers).

b. Feedback about the testing procedure itself: feedback about the testing procedure itself includes information on circumstances and events taking place during the test administration. This may pertain either to activities of the test takers or to circumstances or activities surrounding the test takers. This kind of information is useful in evaluating the degree to which the testing environment supports the test takers in doing their best work.
3. The different sources of feedback

During test administration, feedback will be obtained from test takers, test administrators, test scorers and raters, and test users. Test takers will provide feedback on their perception of and attitudes towards the test and test tasks, and on their performance. Test administrators/proctors will provide feedback on the degree to which the administration procedures have helped the test takers give their best performance. Test scorers and raters will provide feedback on the appropriateness of the test scores and of the rating scales, and on the appropriateness of the time given to score and rate the test tasks. Test users will provide feedback on the usefulness of the selection decisions based on the scores with respect to their particular needs.

4. Amounts of resources involved in obtaining feedback

Bachman and Palmer take the view that the greater care taken in developing a test and the more feedback obtained on its usefulness, the more useful it will be. The process of collecting feedback for assessing usefulness and improving the test should be considered an ongoing activity that is part of every test administration. In order to decide how much and what kind of feedback to obtain in test administration, the usefulness of the feedback must be balanced against the cost (in terms of resources) of obtaining it. In high-stakes tests, gathering feedback might be a large-scale operation. However, in this project, the resources are somehow limited and so the procedures of collecting feedback should be economical since it is important to create a balance between the resources available and the intended use of the test and the methods of obtaining feedback used to check up that use.

5. The method for obtaining feedback

The qualitative instrument that is going to be used to obtain feedback is multiple-choice questionnaires to be given to test takers, test administrators/proctors, test scorers and raters and test users. These economical questionnaires are going to be designed according to the resources at hand.

Test Use

Procedures for Analyzing Test Scores

After scoring the test items, tasks, and subtests and assigning the total score of the test, the next step is to analyze these scores. This analysis includes the following procedures:
1. **Describing test scores:** using descriptive statistics to characterize the quantitative characteristics of the test scores (the histograms of scores distribution, mean, mode, median, range, and standard deviation of the scores).

2. **Item analysis:** for each objective item (facility value and discrimination index using classical item analysis).

3. **Estimating reliability of test scores:** using internal consistency reliability (such as cronbach’s alpha) to estimate the reliability of the objective subtests (reading and listening) and the inter-rater reliability and the reliability of the subscales of the subjective sections (writing and speaking). In addition, “Cronbach’s Alpha if Item Deleted” analysis should be conducted in order to eliminate the items that reduce alpha reliability and make the test more efficient.

4. Calculating item-total correlations: (calculating correlations between the components, and correlations between each component and the total minus that component).

5. **Calculating the standard error of measurement (SEM):** this step is necessary in order to make sure that SEM is maintained at under 5% of the total score. In such proficiency tests, SEM should be kept as a minimum because the smaller the SEM, the more reliable the test and the greater the likelihood that the test taker’s actual score on the test will be close to his or her true score. Since one cannot tell from the reliability coefficient alone how reliable the test is, the SEM should be calculated to provide a better estimate of the test takers’ true scores. If the SEM is more than 5% of the total score, it should be improved by removing and replacing items with low ID values if the test is reused.

6. **Investigating the validity of test use:** for investigating the validity of inferences made from test scores, a number of logical considerations and empirical procedures both qualitative and quantitative should be provided. The qualitative procedures that are relevant to investigating construct validity include answering the questions related to the logical evaluation of construct validity listed in the checklist for evaluating usefulness (Appendix B). On the other hand, there are many quantitative procedures that can be used to investigate construct validity. Two of these procedures can be used for this test considering its limited resources. These are: internal correlations and factor analysis. Correlating the different test components with each other is one way of assessing the construct validity. If two components correlate very highly with each other, they will be measuring the
same trait (or language ability). On the other hand if two components correlate fairly low with each other, they will be measuring different traits or skills, which is the natural purpose of using different components. The second way of assessing construct validity is to conduct a factor analysis by comparing the correlation coefficients of the subtests of the test with each other or with the subtests of another admission test in order to identify the common underlying factor (or factors) shared by these subtests. The findings of the factor analysis are used to ascertain whether or not the subtests measure the same language traits and in order to validate the new test.

**The Logical Evaluation of Usefulness**

Using the information gained from the test scores and the feedback collected from test takers, test administrators, test scorers and raters, and test users, the test developer will answer the questions in ‘The logical evaluation of usefulness’ (Appendix B) to evaluate the six qualities of usefulness. This evaluation will help the researcher assess the plan she developed and used to evaluate the qualities of usefulness in the different stages of test development.

**Post-Test Reports**

Because of the considerable time and energy that have been devoted to preparing and administering this test, it is important for Taiz University to keep records of the admission decisions, the testing procedures, the analyses conducted on test results and the feedback received. It is also important that this information is passed on to the interested audience. The audience is the people working within the institution itself: those in charge of designing future versions of the test and of coordinating related activities. Another obvious audience is the teachers who will be preparing other students to take the admission test in future. There are also other parties who may need information about tests (e.g. administrators in other institutions who may want to use the test and professionals in language testing and related fields who are interested in learning how different testing institutions deal with the classical challenges of validity, reliability and practicality).

Within this institution (Taiz University), the post-test report, which will be written for the institution itself, is intended to have two functions (Alderson, Calpham, & Wall, p. 198). Firstly, the post-test report serves as a historical record of the test because it records how the various testing procedures worked out in practice. Secondly, it serves as a guide for future test development because it contains
recommendations for improving aspects of the current test which did not work out well.

This post-test report will include the relevant statistics (descriptive statistics, item analysis, reliability, and the correlations between the components of the test) and the interpretation of the figures and histograms. It should also summarize what has been learnt from analyzing notes taken during the test administration, examining feedback collected from important participants in the testing process and studying the test takers’ performance.

Archiving

This step constitutes the summit of the test development process. Archiving involves constructing an item bank of test items or tasks so as to facilitate the development of subsequent tests. An item bank is a large collection of previously trialed test items that would be stored on a computer and placed at the disposal of test constructors. The information that would be stored with each item or task includes:

1. a number of identifying criteria such as its content, class level, the testing technique used, and the number of scores assigned to it,
2. the correct response (or responses) and the scoring instructions,
3. its item difficulty or facility and item discrimination indices, and
4. notes on the item (when it was written, when it was used, etc).

This is going to be of great importance to test writers because once they have access to an item bank, they will simply choose from it the items that they need for a test. When they choose all the items they need for the test and provide details such as the test title and rubrics, the computer provides a printed version of the test. Although this practice is not yet used at Taiz University, the researcher, with the help of expertise in the field of compute programming, will start this item banking strategy to be used in order to make the process of test construction easier and the admission tests themselves reliable and more efficient.