The Main Study

Khalid Addamigh, PhD
King Saud University, Saudi Arabia

This chapter comprises the following elements: the research questions, a report on the pilot study, background information on the subjects, an account of the instruments employed, procedure and the test materials used in this study. It also addresses some other interrelated methodological issues about the data collection. The procedure for analysing the data (transcribing, coding, etc.) will be discussed in Chapter V.

IV.1 Research Questions

This study sought to obtain the answers to these two primary questions:

(1) What test-taking strategies do EFL test-takers deploy when tackling MCFGV test items?

(2) If there are different types of test-taking strategies, what strategies occur more than others?

Following the PPP model outlined in Chapter II, the study also sought to answer the following related sub-questions:

(3) Are there differences in the exploitation of test-taking strategies between the high and low TOEFL proficiency testees?

(4) Are there differences in the exploitation of test-taking strategies between the high and low the lexical proficiency testees?

(5) Are there differences between the test-taking strategies used in tackling professionally-made test items and those used for teacher-made tests of the above type?
(6) Could EFL test takers use test-taking strategies to arrive at the correct response even if they did not possess the knowledge intended to be tested in the above kind of test?

(7) If the answer to the previous question is affirmative, which test-taking strategies would lead generally to greater success?

**IV.2 Overview**

This research exploiter a mental process-based instrument to gain insight into the strategies used by university EFL test takers when working on EFL multiple-choice vocabulary test items. The study was based upon a triangulated data collection approach using verbal self-report methods. The intention was to explore the strategies utilised in tackling tests items designed by different test-makers. Thus, the test used in this study for the purpose of identifying strategies was made up of two subtests, professionally-made and teacher-made tests. The sample included subjects of two proficiency levels, as determined by TOEFL, with a view to examining the differences in the strategies used between high and low proficiency students. It study was conducted at two Saudi universities. In each location, eight instruments were employed to obtain the required data. Figure 4.1 provides a summary of the main stages of the data collection process.
### Preparing the Study Test (12 stages) (IV.4.10)

<table>
<thead>
<tr>
<th>Time</th>
<th>Day KSU</th>
<th>Day UQU</th>
</tr>
</thead>
<tbody>
<tr>
<td>2:00</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2:45</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>N.A.</td>
<td>N.A.</td>
<td>N.A.</td>
</tr>
<tr>
<td>0:45</td>
<td>4</td>
<td>15</td>
</tr>
<tr>
<td>0:20</td>
<td>8</td>
<td>N.A.</td>
</tr>
<tr>
<td>0:20</td>
<td>10</td>
<td>22</td>
</tr>
<tr>
<td>0:05</td>
<td>10</td>
<td>22</td>
</tr>
<tr>
<td>ab-</td>
<td>10</td>
<td>22</td>
</tr>
<tr>
<td>out 0:45</td>
<td>11</td>
<td>23</td>
</tr>
<tr>
<td>each</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 4.1: The main stages of data gathering
IV.3  Pilot Study

A pilot study was conducted in the summer of 1999. The following is an account of the purpose, the instrument, the subjects, the procedure and the results of the pilot study.

IV. 3.1  Objective

The aim of the pilot study was to try out the instrument and procedures planned for the main empirical study in order to avoid any mistakes that might occur when embarking on the main data collection. It was, therefore, a testing ground for the intended instrument to check that it was appropriate for the research and, if not, to discover why not and then improve it. Furthermore, as the intended instrument was open-ended, it was thought that the pilot study might reveal whether some of the additional information obtained could aid the construction of another instrument, the questionnaire, which was being considered at this infant stage of the research.

IV. 3.2  Instrument and materials

As discussed in IV.4.1, the introspective verbal report was chosen as the principal method in the main study. Accordingly, it was the method by which data was collected in the pilot study. There were two test item formats examined: gap-filling through multiple-choice and matching A with B. In the pilot study, the aim was to investigate TTS for the two test item formats that were constructed by the teachers alone, as the intention at that time was to examine TMT only. Each test format was represented by five items derived from the Saudi teacher-made achievement tests that had been collected for the preliminary study. The items were selected to be of a suitable level of difficulty for the proficiency levels of the test-takers.

IV. 3.3  Pilot subjects

Four Saudi students took part in this study. Their ages ranged from nineteen to twenty-eight. Three were studying English in order to progress to postgraduate studies in the UK and one wanted to improve his language proficiency level for general purposes.
They had all studied English whilst they were at school in Saudi Arabia and were studying English at the Colchester English Study Centre in England at the time of their participation in this investigation. They had been in the UK for periods ranging between one and five months. According to the centre’s placement test, their language proficiency levels were elementary in one case, intermediate in two cases and upper-intermediate in the fourth. The ESL pilot subjects were not exactly like the EFL subjects of the main study, but they were close enough to enable useful information to be obtained.

**IV. 3.4 Procedure**

A short practice session was provided for the subjects to acquaint them with the data-gathering situation. An example by the researcher was presented and they were given some practice on how to ‘think-aloud’ when they met with unknown lexical items in a reading comprehension text. The materials used were taken from unstudied units of their coursebooks, *Life Lines* (Hutchinson, no date) and *Headway* (Soars, no date). Before taking the test, which had no time limit for completion, oral and written instructions in the L1 were provided to encourage the subjects to take the test as if they were under normal examination conditions and to speak out their thoughts, in their mother tongue or in English, whilst tackling the test. The subjects, who took the test individually, were tape-recorded and the protocols were then transcribed but not coded or analysed.

**IV. 3.5 Implications for the main study**

In general, the data obtained was not satisfactory. The audiotapes showed that two of the subjects provided little data, one protocol was acceptable and only one of the protocols could be considered as sufficient. Later, the three subjects were asked why they did not verbalise sufficient data but their replies were not clear. It appeared, however, that the instructions and materials for the think-aloud session were poor, resulting in an unsatisfactory outcome. Furthermore, the subjects were acquaintances
of the researcher and may have felt somewhat self-conscious about verbalising their thoughts. Accordingly, the pilot study indicated that in order to overcome the first shortcoming, the three improvements below should be considered in the main study:

(a) Greater attention needed to be paid to the introspective training session in the main study.

(b) It would be more appropriate to avoid gathering the data on an individual basis, by using a language laboratory for example.

(c) It may be necessary to carry out immediate retrospective verbal data gathering and/or interviews along with the introspective method to ensure more feedback on the strategies used.

With regard to the second problem, ‘shyness’ would not affect the verbal protocols for the main study in the same way as it did for the pilot study, or at least the effect would be less, as the intended subjects would not know the researcher.

IV.4 Main Study Subjects
Thirty-two Saudi EFL undergraduate students took part in the main study. They were randomly selected from two distinct groups of students enrolled in two universities. The two groups, which were equally represented in this study, were selected to be of a similar range of levels of proficiency (see section IV. 4.7) and were freshmen. The first group (KSU) was chosen from King Saud University. The group was made up of university students attending an intensive English-for-academic-purposes course at the EFL unit at the College of Languages and Translation. They were taking pre-sessional courses in English in order to qualify to study at various medical faculties at KSU where they would mainly be studying in English. The second group (UQU) was chosen from Umm al-Qura University. This group was made up of full-time students majoring in English in the English Language Department.
Both universities are government-run public universities. The former is located in Riyadh, the capital of Saudi Arabia, in the middle of the country, and the second is in Makkah, the religious capital of the Islamic world, about nine hundred kilometres west of Riyadh. These two universities were chosen because of their different reputations. King Saud University is one of the best educational establishments in Saudi Arabia and attracts students from all over the country. Umm al-Qura was chosen to provide a balance as it is a relatively small university among the eight universities in Saudi Arabia and it has the reputation of being of average standard. The study, therefore, could be said to cover a representative sample of Saudi students, which helps to validate generalisation of the results of this study.

Since all the subjects came from similar backgrounds the data obtained from the two groups was treated as coming from one group of thirty-two subjects. The intention in conducting the study on with two groups was to gain the following key advantages:

(a) To create more opportunities of conducting immediate retrospective interviews.

(b) To provide a better representation of the population.

(c) With one group there would be one test of fifteen items whereas with two groups two dissimilar tests would be taken, each consisting of fifteen items. This is because each group had different teachers so the TMT items would be different, resulting in thirty dissimilar items. Two groups would give a more widely representative sample of the test items examined.

The subjects, who participated voluntarily in the study, came from different towns in the country and appeared to be representative of the typical socio-economic background of the country. They were aged between nineteen and twenty-one. They were all male, which meant that the gender variable was held constant on one value. For the purposes of this study, it was not possible for the researcher to access and subsequently consider female university students from Saudi as in Saudi Arabia men and women are educated
separately. All the subjects were native speakers of Arabic and English was their sole
foreign language. As regards the Saudi students’ attitude towards learning English, it is
generally low. Hafseth, Hickman and Watson (1993) state in their review of the Saudi
English syllabus that most students regard English as a school subject they were
required to pass, rather than as a useful language to learn. In their view, it is a
component in the process of completing their schooling. Students’ attitudes could
change, however, at university level, especially if they choose to major in English or
need English for their future studies. Eighteen of our subjects, of both high and low
proficiency levels, were asked a closed question about their attitude toward learning
English:

‘How important is it for you to become proficient in English?

a) very                         b) quite                   c) not at all’

Thirteen replied “very”, five replied “quite”, and none replied “not at all”. This
indicated that, at the time when the study was conducted, the subjects had a good
attitude towards learning English.

Since the general educational system in Saudi Arabia is the same all over the country,
the subjects had more or less an identical history of schooling in the English language.
All were treated as beginners in English when they began learning it as a compulsory
subject in grade one at middle school at the age of twelve. They received three years
instruction in English at middle school and a further three years at secondary school.
Most of the subjects had obtained high scores in their English examinations for the
secondary certificate, as we inferred from the fact that they had gained admission into
colleges, where admission is based not only on interest in the subject but also on scores
in the Secondary Leaving Certificate, which includes an English examination.
Regarding their experience of English vocabulary test items, the preliminary study
(Chapter III) showed that the contexts varied but the most common type was the
multiple-choice gap-filling format. The subjects were asked directly about this format
and they affirmed that they were familiar with it. In addition, they confirmed that they
had not taken any training courses on TTS.

IV.5 Instruments and Procedures

Eight instruments were used to elicit the data required for the present study.

(a) Three verbal-based instruments were used to provide the data on TTS. They were:
(1) introspective self-verbal reports (VP), (2) immediate retrospective VP and (3)
immediate and later retrospective interviews.

(b) Five paper-based instruments were used to gather interrelated data. They were:
(4) a TTS checklist, (5) a general language proficiency test, (6) a vocabulary size
test, (7) an item difficulty-rating instrument and (8) the multiple-choice
vocabulary test (the study test) used as stimulus for the verbal reports.

The applications of these instruments were identical in both KSU and UQU. Although
there were a number of instruments used, the application of there went very smoothly at
both universities. This was because:

(a) The researcher has been a member of staff at King Saud University for about
thirteen years. This ensured co-operative contacts at both universities, which
allowed the study to take place on a personal basis where necessary.

(b) There was considerable support from the official channels via which the researcher
carried out the fieldwork. For example, before the researcher went to collect the
data, a letter from the Saudi Cultural Attaché in the UK had been faxed to the
Dean of the Faculty of Languages and Translation at KSU confirming the need for
the research. Consequently, the laboratory technicians co-operated fully. The
researcher also presented the Vice Chancellor of Postgraduate Studies and
Academic Research at UQU with a letter from his University (KSU) that
requested him to facilitate the progress of this research. The Vice Chancellor
contacted personally and in writing those who would be involved: the dean of the
Faculty of Humanities, the chairman of the English Language Department and the
director of the EFL Unit. Thus, the work went smoothly there as well.
The administration of the instruments took approximately seven and a half hours of the subjects’ time. This can be broken down as follows:

1- Approximately two hours and forty-five minutes for the general language proficiency test.

2- One hour for the English lexical proficiency test and the instructions. (The actual test took only forty-five minutes).

3- Two hours for the think-aloud training session.

4- Almost one hour for the study test, with the introspective and retrospective reports, as well as the instructions needed.

5- Approximately forty-five minutes for each interview.

IV. 5.1 Triangulation

To increase the validity of the results obtained, the method utilised for data collection in this research was based on triangulation. Cohen and Manion (1994) identify triangulation as the use of two or more means of data collection focused on the same target variables. Consequently, introspective and immediate retrospective VP and semi-structured interview were used as the instruments through which the strategy data was gathered. These open-ended research tools were the most appropriate means of delving into the uncharted territory of this research. As discussed in Chapter II.8, the use of other instruments, such as a questionnaire, would have required previously extracted data relevant to the area investigated, to act as a guideline for the content of such an instrument. However, to the best of the researcher’s knowledge, this study is the first in that particular area of research and, therefore, sufficient and adequate data could not be obtained for the construction of such an instrument. Furthermore, although a questionnaire has its own advantages, it may not be an appropriate instrument for gathering data involving the short-term memory of mental processes because of its lack of immediacy. Oxford (1996) emphasises that the questionnaire is not a useful measure for identifying the particular strategies used for a specific language task at a specific
time. Other scholars admit that, in contrast to concurrent verbal protocols, questionnaires are more likely to evoke the learners' attitudes and beliefs about what they generally do, or have done, rather than what they actually do whilst being involved in a specific language task (e.g. McDonough, 1995; Cohen, 1996).

Figure 4.2: The triangulation scheme for data collection

It should be emphasised that the retrospective reports and interviews were employed as complementary or supportive evidence for the introspective data, in order to achieve a clearer picture of anything that was incomprehensible when the students were involved in the test. The feedback from these two instruments, therefore, was neither coded nor analysed independently. The data from each of them was analysed and coded with the introspective think-aloud protocols, to add more validity.
Because both the introspective and retrospective VP were new to the intended subjects, a two-hour training session had to be carried out to acquaint them with this method. The following section will give details about that phase.

IV. 5.2 VP training session

There were two reasons for deciding that extra attention should be paid to the VP training session: (a) the theoretical background mentioned in Chapter II.8.1.5, e.g., a certain amount of training improves the quality of data (McDonough, 1995) and (b) the experience of insufficient data collection in the pilot study (for more details see IV.2.5).

The administration of the training sessions was similar for both groups of subjects in the two universities. It was the first meeting with the subjects, so the researcher acquainted them straight away with the objective of the study and the importance of the participants’ roles in it. The students were then informed that if they participated in the study they would need to meet the researcher five times for introspective report training, TOEFL, a vocabulary level test, the study test and the interview. They were also told that if they wished to withdraw from the study they were free to do so at any time without repercussions. However, none of them withdrew.

In the training session, care was taken to avoid training the subjects in TTS themselves. Therefore, neither multiple-choice test items nor tasks on vocabulary were chosen to be the materials for that session. This decreased the influence of the training models on the outcome of the study, and, in turn, increased the internal validity of the study. The practice exploited two tasks:

(a) Presenting a picture where an item was missing and a star replaced it. The task was for the students their voice thoughts while trying to guess the absent object, using the other elements in the picture as clues to help them find the missing item (one of the pictures is given in Appendix A.1). This task was similar to that used in the data collection, where there was a stem with a missing word.
(b) Giving an imaginary situation and presenting pictures with similar options. For example, a picture of four people was given (see Appendix A.2) and the students had to imagine that one of the four people was waiting for a taxi. They were asked to say what they were thinking about when they tried to identify the person. This was similar to the problem of choosing from unknown alternatives.

The training began by drawing the participants’ attention to the three stages of learning events: input, process and outcome. We explained that our focus was on the middle one which could be explored by self-verbal report. They acknowledged that this method is becoming widely used in academic research and it simply requires them to say their thoughts. Since, if models were presented by the facilitator of the session (the researcher) the candidates might mimic them, the subjects were left to introspect as they liked. In that session some of the above tasks were distributed and the practice went through three stages in the order: all, pairs and then individuals, as follows:

1- A number of the subjects participated in saying what they were thinking whilst they were trying to tackle the tasks. At the same time, the other subjects were either participating or listening to their peers’ thoughts.

2- The subjects worked in pairs to think-aloud to each other while they were tackling other problems.

3- All the subjects worked individually, to voice their thoughts while each was processing a different task.

The result of the introspective reports in the study test showed, generally speaking, that the subjects provided satisfactory data. This may be because the training session was successful and/or because the subjects’ motivation for taking part in the study was high. In short, the training tasks appear to have been appropriate for both the study purpose and the participants’ interest.
IV. 5.3 Introspective protocol

As mentioned in Chapter II.8.1, among the methods used in exploring the learners’ strategies introspective VP has been regarded as a valuable tool in eliciting data that is rich in quantity and quality. This led to the growth in its popularity amongst researchers (Pressley and Afflerbach 1995; Green 1998). Ericsson and Simon (1993) claim that the most useful instrument that can be used to obtain mental data is the verbal protocol via immediate think-aloud, because it concurrently reflects mental processes. Consequently, self-verbal introspective report was utilised as the main means to gather data relevant to the strategies targeted in this study. As suggested in the pilot study (IV.2.5), this instrument was administered, in the two test-taking sessions, in language laboratories at both universities. The subjects were familiar with the location (details about the location are in IV.4.9.6).

Written and verbal instructions were provided regarding the introspective reports. The first sheet of each of the two tasks presented to KSU and UQU students contained, in Arabic, general information as follows:

(a) A short statement emphasised the importance of the students’ roles in the study, as the whole project depended on their active participation in this task by vocalising their thoughts.

(b) This was followed by three instructions that guided the think-aloud method based on the following:

- **I am interested in whatever comes to your mind** no matter whether it is a good or less than a good idea or a question. (Ericsson and Simon 1996).
- **Do not rationalise or plan what to say**, but rather let us hear your thoughts (Green, 1998).
- **Be bold!** Everything on your mind is important to us, so say it all even when you change your mind.

(c) A request for some information about the test takers.
Also, as suggested in II.8.1.4, written prompts were given in the form of an Arabic phrase that says, “Do not forget to say everything on your mind”, which was combined with almost every item. Anderson and Simon (1996) claim that a reminder of the type that asks “What are you thinking about?” is likely to elicit a self-observation.

Furthermore, immediately before the tests took place, the subjects were given verbally the information and instructions below:

- They were encouraged to take the test as if they were under normal examination conditions and reminded of the prizes for the highest scores achieved (details about the prize are in IV.4.10.4).
- They were requested to verbalise everything that came into their minds when they performed the test. They were also reminded of the prizes that would be awarded to the students who verbalised their thoughts the best.
- They were informed that any information they provided would be used for academic purposes only and, even if it was be mentioned in the research, it would be anonymous.
- They were reminded that they were free to voice their thoughts in their L₁ or English or to shift between the two languages.
- If they wanted to ask the examiner a question during the test they would be able to do so by pressing the tutor button in the booth.
- The subjects were told about the time available for the test. If they finished before the time was over, they would need to turn the tape recorder off and wait in their seats until the others had finished.
- They were then asked if they had any comments, a question, or anything that needed to be clarified.
• Finally, they were requested to write their names on the audiotapes, making sure that side ‘A’ was up when inserting the tapes into the machines, turn the tape recorder on and wear the headphones.

After there instructions, the study test was handed out to the subjects and they began the test, voicing their thoughts. Thirty-two subjects were involved in the study and the test involved two types of different subtests. That means that there were sixty-four reports. For more details of the procedure see section IV.5.

IV. 5.4 Immediate retrospective protocol

The second data collection instrument was immediate self-verbal retrospective reports. In view of the fact that the data obtained in the pilot study via introspective oral report was insufficient (IV.2.5), retrospective reporting was used in the main study in addition to the introspective reporting to review what had taken place. This instrument was employed to ensure that subjects who may have done the test silently would say something about their mental process in the second phase of verbalisation. Ericsson and Simon (1993) and Green (1998) claim that valuable information about cognitive processes can be provided by both introspective and retrospective reports. In their view, the former can indicate the information that is provided by the subjects as they perform mental tasks, while the latter reflects the subsequent recall of those sequences. The objectives of using the retrospective data gathering were:

(a) To further validate the elicited strategy-related information.
(b) To gain a better understanding of what was unclear in the first phase.
(c) To obtain more data in order to tackle, even if only partly, the problem of data incompleteness that might occur in the introspective reports (more details about incompleteness are in II.8.1.1).

All of the subjects who supplied the first VP also gave retrospective reports. As soon as the test time was over, the test takers were instructed verbally to go through their
choices, without changing their answers, and say why they had chosen each response. This stage can be considered as the stage of immediate retrospective reports that Cohen (1998) classifies as a 'self-observation' method. When all the subjects had gone through the retrospective phase, they were requested to leave the laboratory and leave the test protocols behind in the booths. This would allow the audiotapes to be matched with their corresponding written test protocols, as some subjects might have forgotten to write their names on the audiotapes.

**IV. 5.5 Retrospective interviews**

The semi-structured retrospective interview was used in order to obtain additional data about the same task and to add more validity to the information obtained. Since the interview was conducted to detect the strategies used by the subjects when they were involved in the test, the interviewer investigated the data elicited from the introspective phase, as it immediately reported the subjects’ behaviours. In the interviews the subjects were questioned individually by the researcher about the information they provided in their recordings. The interviewer’s questions, therefore, were not prepared in advanced but were directly related to specific actions the subjects’ displayed in their test. Thus, according to Nunan’s (1992) classification, the interviews can be labelled as semi-structured. The interviews examined both protocols:

(a) The introspective report, as the audiotapes were run through with each subject, to investigate unclear thoughts, e.g., “What do you mean by this?”; “What were you thinking about here?”.

(b) The test protocol, as the subjects were asked to explain any noticeable marks on the test sheets themselves that were not spotted by the former method, such as any underlined or circled words.

The original intention was to conduct interviews with all of the subjects to whom the first instrument was administered. However, as the subjects took part voluntarily and as the other phases of study took more than six hours (details are in IV.4), some of the
subjects found it difficult to spare the additional time for the interview. Interviews were conducted, though, with an almost equal number of high and low proficiency subjects from both KSU and UQU. In the introspective training session (which was the first meeting between the subjects and the researcher), a timetable was distributed to the subjects and they were asked to indicate the most convenient time for them to have the interview. This diary included times for meetings immediately after the think-aloud test-taking and in the two days afterwards, as well as the subjects’ contact numbers, so that they could be contacted in case the researcher needed to change the arrangements. Five subjects from KSU and three subjects from UQU were interviewed immediately after taking the test. This can be considered as immediate retrospection. In addition, four subjects from KSU and four from UQU were interviewed the following day, and two subjects from KSU were interviewed on the third day. This means there were eighteen either immediate or later retrospective interviews which lasted on average forty-five minutes. To enable the participants to express themselves completely and fluently, the interviews were conducted in their native language. They interviews were carried out with each subject individually and took place in the language laboratory for the KSU group and in an office in the subjects’ department for the UQU group.

In conducting the interviews the researcher used a previously prepared instrument, which the following section will discuss.

**IV. 5.6 Checklist**

The fourth instrument deployed in the study was a checklist, Appendix B, that covered possible strategies the test-takers might utilise for the test. It was developed to be used by the researcher only in the interview phase, to enable him to note down the number of strategies used, if they were on the list, instead of writing down the subjects’ spontaneous comments on their reports. The objective of this instrument, therefore, was to help the interviews to be carried out more quickly. The strategies incorporated in the checklist were derived from the questionnaire that was considered, at an infant stage of
this study, as a potential means of data collection and were constructed in the light of the following:

(a) The data obtained from the pilot study;

(b) The literature on the strategies found in other fields of TTS (more details can be found in Chapter II.9);

(c) Instructions about constructing test items taken from works on testing.

IV. 5.7 TOEFL

This study attempted to consider the differences in TTS between high and low proficiency students. The Test of English as a Foreign Language (TOEFL) was used to demonstrate the subjects’ relative proficiency levels in English. The Educational Testing Service (ETS) gave permission to the researcher to use the TOEFL for this study and recommended ‘TOEFL Practice Tests, 1995’, which included materials from actual tests given at a worldwide test administration (ETS, 1995). The TOEFL was chosen not only because of its popularity but also because other examinations have an assessment of speaking skills. Such assessment would need to be conducted by a professional examiner and would prove difficult to arrange. Ignoring this oral section or having it assessed by someone other than a professional examiner, would adversely affect the validity of the test. TOEFL was, therefore, easier to use, especially because only a relative measure of proficiency was required and oral proficiency was not relevant to our study.

The number of students who took the TOEFL was larger than the study required. This was in order to create a gap between the high and low proficiency subjects. The highest and lowest proficiency students in each group were chosen to be our sample, and the students of average proficiency were withdrawn, to provide a clear gap between the high and low proficiency students. The size of the gap was more or less one-third of the number of students that took the TOEFL. Before administering the TOEFL, as almost none of the participants had taken the test before, information about how to
answer each section of the test was provided. The test was administered to the two
groups in language labs at the subjects’ universities. The participants were pleased to be
given an opportunity to find out their level of attainment on the TOEFL. It appeared that
it was a useful experience for them (the result is in Chapter VI).

IV. 5.8 Nation’s vocabulary levels test

This instrument was used to establish the subjects’ levels of English lexical proficiency
in order to find out if there is a relationship between this dimension of language
proficiency and the subjects’ use of strategies in tackling vocabulary tests. It must be
pointed out that although other lexical aspects might be relevant, such as depth of
vocabulary knowledge and lexical organisation, only the impact of lexical size was
examined in this study because this dimension appeared to be the ground base for the
other lexical aspects in the L2. Qian (2002) concludes that depth and breadth of
vocabulary knowledge are closely and positively associated with each other. Meara
(cited by Schmitt, 1997: 301) suggests the intuitively reasonable proposition that people
with a wide vocabulary tend to demonstrate a deeper knowledge of individual words
than those with limited vocabulary. Hence, Nation’s Vocabulary Levels Test (NVLT)
(Nation, 1990) was used to establish the subjects’ English lexical proficiency. The other
available instrument that could have been used for this purpose was Meara’s (1989)
Test. Nation’s measure was chosen because it essentially tests receptive knowledge of
the meanings of words, which is the type measured by our test format, and also because
its format considerably narrows down the chances of guessing correctly. It was also
suitable because it is of a different format from our test (2/5CI, see Chapter III for more
details). All the participants who took TOEFL took this test too. The aim was not to
form a gap between the subjects with high and low vocabulary proficiency, as with the
TOEFL test. It was given to all the participants, because it was not known at this stage
who would be withdrawn. For the KSU group, NVLT was conducted two days after the
TOEFL. The UQU group took the test a week later (the results are given in Chapter
VI).
IV. 5.9  Item-difficulty rating instrument

The seventh instrument was constructed to estimate the level of difficulty of the items in the study test, as perceived by language teachers. It was a straightforward rating instrument in two versions, one for the KSU test and the other for the UQU test (Appendixes C.1 and C.2). The objectives of establishing the level of difficulty of the tests were as follows:

(a) To make a valid comparison between the set of strategies used on each of the two types of test stimuli (i.e. PMT and TMT) it had to be ascertained that they were of a similar level of difficulty.

(b) It was necessary to make sure that the test items used were neither so easy, that the answer would only need simple direct lexical retrieval processes from the ML, nor so difficult that they would be beyond the test takers’ abilities.

The test items of the two types used in the study test were systematically mixed: the first item was PMT and the second was TMT and so on. Each item was followed by a scale of five degrees to identify the level of difficulty of the item. This was followed by spaces for any further comments on the stems or the distractors as the instructions for this rating instrument encouraged the participants to pay particular attention to the stems and distractors.

Level of difficulty is, of course, a relative concept and depends on the subject’s proficiency level. The rating was carried out, therefore, by university EFL language instructors who were aware of the test takers’ levels of English language proficiency, as they were from the subjects’ educational department/unit. However, those who carried out the rating were different from those who constructed the tests. Four of the KSU language instructors rated items of the first version of this instrument, and four language instructors from the UQU rated the second one (the result is given in Chapter VI).
IV. 5.10 The study test

The eighth instrument was the study test, the vocabulary test used for the VP to detect the TTS utilised by Saudi EFL students. The test format of the study test was based on the results of the preliminary study (see Chapter III for details). There were two dissimilar versions of the study test for two distinct groups of test-takers, one for KSU and the other for UQU. The present study attempted to explore the TTS for tackling two types of tests made by different types of test makers: (a) professionals and (b) teachers. Consequently, the study test was made up of two subtests: (a) the PMT and (b) the TMT. The KSU test version comprised sixteen items (eight PMT and eight TMT) and the UQU version included fourteen items (seven PMT and seven TMT). This means that fifteen test items represented each type of test maker (i.e. PMT and TMT). Each task specification for the subjects was of three sheets long. The first sheet, which was in Arabic (English translation in Appendix C.3), gave general information about introspective reporting during the test (more details are in IV.4.3). The other two sheets contained the two types of tests, one involving PMT items and the other TMT items (Appendixes D.1 and D.2). The PMT items used in the two tests were designed either by the Local Examinations Syndicate at the University of Cambridge in the UK, or by the Testing and Certification Division of the English Language Institute, University of Michigan, USA. The teacher-made test items were constructed by five university EFL language teachers of four nationalities in KSU and UQU. Three of the teacher had a Master’s in TESL, one had a Master’s in Translation and the fifth had a PhD in TESL. They were of different nationalities: two Algerian, one Nigerian, one Saudi Arabian and one Pakistani. Their ages ranged between thirty and fifty years.

IV.5.10.1 Guiding principles of the study test

The question that posed when considering the construction of a test for the current purpose was: on what should the choice and development be based? To answer this question, we needed to consider the intended subjects’ situations. Their exposure to English was effectively confined to the English syllabi used because English
is not used by the vast majority of [Saudi] students in their daily lives outside English classes and it belongs to a foreign culture that they have had little or no exposure to (Hafseth, Hickman and Watson, 1993: 2).

The PMT and TMT items, therefore, were constructed or selected by using a syllabus-based method, making our test more one of achievement than of proficiency, similar to the tests usually taken by Saudi students. In other words, the study test items were based on the English programme being taught to the subjects when the empirical study was being carried out. Relying on other means, such as word frequency use, in order to have two sets of lexical items of similar levels of difficulty might not have been a valid method. Existing frequency lists, for instance, had been designed in an environment removed from the intended subjects’ backgrounds. One of the differences is that ‘more frequent words tend to be less formal and less frequent words tend to be more formal’ (Schmitt 2000, p.6). The subjects of this study, were likely to know more formal words than informal ones, as their exposure to English had been confined to their educational environment.

IV.5.10.2 Choosing PMT and TMT items

Another aim in the development of the study test was to match the two sample sets of test items (PMT and TMT) as far as possible, especially with regard to difficulty. Manipulating the stems or options of the PMT/TMT items for the test, in order to obtain equivalent sets, would have resulted in one or both of the test types losing some of their natural characteristics which make up the dissimilarities between the two types. For example, if it were found that some TMT items included distractors that were easier to identify as incorrect responses than the ones in the PMT, and these distractors were then changed so that they were less obvious, the new items would no longer be representative of an actual teacher-made test. Thus, an attempt was made to make the two types have corresponding levels of difficulty without modifying the items made by the two types of test makers. Accordingly, in order to choose equivalent test items, a
greater emphasis was placed on the target lexical items, as they are the essence of any vocabulary test item.

To obtain this objective, the twelve courses of listed action below were followed:

1- Since the study concerned a particular language task (i.e. vocabulary) and a specific test item format, restrictions were placed on the sources of the intended tests, such as:
   (a) The items should test vocabulary knowledge.
   (b) The test item format should have multiple-choices for sentence gap filling.

2- Various PMTs were searched for and examined to choose the one/s that met the above criteria. It was found that the Cambridge EPT and Michigan examinations (EPT, ECCE and MTELP) met the above conditions (samples are given in Appendices G.1/2/3). They are, of course, proficiency tests for EFL learners of any L1, so needed further procedures to yield suitable items for our study.

3- The researcher contacted the two establishments via official means to obtain as many diverse examinations as possible. He successfully gained an adequate supply of tests. Some of the Michigan examinations (MTELP) were quite old, so only the more recent forms (Q and later forms) were considered.

4- Seventeen different authentic examinations were obtained, yielding more than five hundred multiple-choice test items aimed at testing vocabulary meaning knowledge.

5- The target words of these items were placed according to their word class in a list of four hundred and seventy-two lexical items (Appendix E). The number
of lexical items listed was less than that of the vocabulary items obtained. Some of the Michigan test items were omitted as they involved identifying a synonym for an underlined word from multiple-choices rather than filling in the gap.

6- When a word had a particular meaning in the test context but might have a different meaning when isolated, the word was accompanied by a collocation to specify the meaning in the test. For instance the word meet was put in the list as: meet (meet a standard).

7- The subjects’ language teachers at both universities were asked to identify from the list of 472 lexical items only 14 which the students had already learned as part of their current university course, regardless of whether or not these lexical items might have been taught earlier at school. In other words, the criterion was that these lexical items were part of the knowledge that could be measured in any achievement assessment in their current course.

8- The requested number of lexical items, 14, were identified for UQU subjects, and 16 lexical items were identified for KSU subjects (two more words were identified). The researcher decided, therefore, to give this group 16 test items.

9- The lexical items identified for each group were put into two distinct lists (sixteen for KSU and fourteen for UQU). Subsequently, every second word of each list was put randomly into two other sub-lists (eight for KSU and seven for UQU).

10- Each of the three EFL language instructors at KSU and the two at UQU were given the relevant sub-lists. They were asked to construct, in their usual manner, items in the filling-in-the-gap sentence-based format using the lexical items in
the sub-list as the target words. This resulted in 24 items being constructed by KSU language instructors and 14 by those at UQU. The language teachers were asked whether they typically used this format and the responses were affirmative.

11- Eight different items from the ones constructed by KSU language instructors and seven from the UQU language instructors’ items were randomly chosen to represent TMT. The instructors knew in advance that not all the items they constructed would be used. In their item construction, they used the same word class of words as in the list, except for one who used the adjective *specific* instead of the listed adverb, *specifically*.

12- Finally, the test items in the professionally-made examinations that tested the remaining eight and seven lexical items on the other two sub-lists were picked out to represent PMT.

Items made by the two types of test-makers, PMT and TMT, were:

(a) selected in an equally random way to achieve equivalency of items tested
(b) legitimately able to be presented as an achievement test. This would be consistent with the analysis in II.1, which showed that the type of test most encountered by Saudi students is, solely, the achievement test.

The average level of the subjects’ language proficiency, as shown by TOEFL, can be considered as intermediate. In contrast, the vocabulary test items used in this study were derived from various tests, some of which were designed to examine more advanced students. Nevertheless, these vocabulary items were picked and used in the present study because they demonstrated the specific knowledge that had been required to be studied. In other words, they were no longer used to test the general level of language proficiency. However, the lexical items identified by the teachers, as occurring in the
subjects’ course materials from the advanced level tests, were fewer than the ones from the intermediate level and placement tests.

**IV.5.10.3 The level of difficulty of PMT and TMT**

Since the study aimed to investigate whether or not there are differences between the strategies used in tackling PMT and TMT, the two types of item had to be carefully chosen to avoid any discrepancies in level of difficulty. If they were not of a corresponding level of difficulty, the difference, if there was any, might occur because of the difficulty variable and not because of the difference in design by different test makers. In this study, the researcher intended neither to construct the items himself nor modify them. Consequently, various means were used to determine the level of difficulty of the two test types. These were:

(a) The syllabus-based means: the method through which the two sets of target word and items were chosen (IV.4.10.2). this was a solid base for obtaining two tests of similar levels of difficulty.

(b) The test-taker-based means: the subjects’ scores in the two tests were examined to determine the difficulty level from the perspective of the candidates’ performance (the results are given in Chapter VI).

(c) The teachers’-views-based means: this was via the item-difficulty rating instrument, see IV.4.8 for details (the results are in Chapter VI).

This triangulation approach was used to add more validity to the results obtained. However, we have to admit that only (a) was used in making the study test and (b) and (c) were analysed with the results and used as a check. That was because it was not possible to use (c) or (b) at the time of the test construction.
**IV.5.10.4 Motivation for taking the test**

Despite the ‘pseudo’ of this investigation, every possible effort was made to collect data under real examination conditions. To make the scores of the study test as important for the test takers as they would be in an authentic test, an application was presented to the EFL Unit at KSU to consider the results of the research test as part of the subjects’ course achievement marks. The Unit could have, for instance, omitted a vocabulary section in the final test and used the results of the study test instead. The application was unsuccessful. The course director decided not to accept the proposal as there would be discrimination in the assessment between the groups, particularly as the final test would be issued in one version for all the eight groups that studied the programme. However, even though this goal was not achieved, positive wash-back was gained on the study test conditions. This was because, when the subjects actually took the test, they were aware of the negotiations and the fact that the test result might have been considered as part of their final marks. In order to encourage the subjects to participate adequately, five prizes were offered in each location, two for those who scored the two highest results and three prizes for the three students who verbalised their thoughts the best. The prizes were valuable, stylish wall-clocks with the King Saud University logo printed on them. In the first meeting with the subjects it was emphasised clearly that they were free to take part or not in the study. It is worth mentioning, however, that the participants were pleased to take part, took it seriously and were very co-operative.

**IV.5.10.5 Date of test administration**

The data collection phases of this study were performed during the subjects’ class time in the first term of the academic year 2000/2001. The study test was taken towards the end of the term, exactly three days before the last day of the actual study programme for KSU and a week before the end of the subjects’ course in UQU. At KSU, the subjects took the test on Saturday and then normal classes took place on Sunday (a working day there). On Monday there was a second midterm quiz, then a break for Ramadan and Eid, followed by the final exam. At UQU the subjects took the test on Tuesday and the
course finished on the Wednesday of the following week. Once again the final exam followed a break for Ramadan and Eid.

IV.5.10.6 Ordering

The study test was made up of two subtests, i.e. PMT and TMT. To maintain the integrity of each subtest, in case this affected the responses, items of the two types were not intermingled but two test forms were used for each university. Form ‘A’ started with the professionally-made items, and form ‘B’ started with the teacher-made items. The subjects were not told that there were two subtests. They were divided into two groups made up of students from both the high and low proficiency groups. The first group was tested with form ‘A’ and the second group with form ‘B’ (see Figure 4.3). In this way any influence of the subtests’ order could be avoided.

![Diagram showing two versions of the study test to control the influence of order]

**Figure 4.3: Two versions of the study test to control the influence of order**

IV.5.10.7 Location

The two test-taking sessions took place in two modern language laboratories in order to enable the verbal self-report to be audiotaped. The subjects were familiar with the location as their listening-comprehension lessons were usually conducted there. Some
of their tests took place there because the rooms were large and the seats were relatively well-separated. Twenty candidates sat for the tests and did the introspective phase in each session. Each lab was equipped with forty-eight booths, and each booth had three partitions. Once the headphones had been put on, it was not possible to hear anything from the next booth. However, two actions were taken to avoid any possibility that a test-taker might be tempted to copy the strategies or answers of another: (a) adjacent test takers were given different forms and (b) every second booth was left empty during the test. Figure 4.4 illustrates the test location. The two labs were of similar design except that the control board was at the front in UQU’s laboratory.

Figure 4.4: The laboratory where the subjects took the test & thought aloud
The day before the study test, the labs were checked and prepared by a technician and the researcher. The tape recorder heads were cleaned and the tapes were put in the booths. The booths were then labelled with the subjects’ names, to facilitate the work on the day of the study test and to avoid any errors that might occur as there were two forms of the test (A and B) and two groups of subjects, high and low proficiency, and a similar number of each form had to be taken by each group. The laboratory made the study test atmosphere close to the subjects’ ordinary test-taking situations.

IV.5.10.8 Time allocation

Bachman (1990) claims that difficulty created by speed or time pressure obviously influences a test, as it may lead to guessing or compensatory test-taking strategies. Thus, in order to preserve the timing integrity of the testing condition, a preliminary investigation was carried out to estimate the time needed for the test-taking session. The test time was estimated by a trial conducted on a sample of students who were taking a similar intensive English language course at KSU. The sample group, which was distinct from the group of intended subjects, consisted of six students of both high and low proficiency levels, equally represented. The opinion of one of their teachers and the results of a recent reading quiz were taken into account when deciding their proficiency level. Ideally, it would have been more valid and reliable if the sample’s language proficiency had been determined by a standardised test identical to that taken by the main subjects. However, this was not possible, due to time restrictions and the availability of the sample. The test used in this preliminary exploration was the same test that was designed for the first procedure at KSU, which included sixteen items (Appendix D.1). The students in the sample group were instructed to do the test as they usually did, to finish it as soon as possible and to obtain the highest marks they could. When a candidate finished, the researcher noted the time and rounded it off to the nearest minute. The result is summarised in the table below:
If the average time for completing the preliminary test (15.17 minutes) had been made the actual time allocated for the final test, it would mean that approximately two thirds of the test-takers might not have completed the test within the allocated time. Bachman (1990) states that in a power test, such as ours, ‘sufficient time is allocated for all, or the vast majority of the test takers to attempt every item’ (Bachman, 1990:123). Accordingly, the average time needed for the lower proficiency sample group to do the test (17.5 minutes) was used to estimate the time to be allowed for the main study, in order for the vast majority of the test takers to have time to attempt every item.

As stated in Chapter II.8.1.3, extra time is needed to complete the task with thinking aloud, as the subjects are required to perform two tasks at once (Afflerbach and Johnston, 1984; Fletcher, 1986). Thus, extra time was given to allow for the verbalising of thoughts. As far as the present writer is aware, time estimation for the think-aloud has not been examined, so it was estimated to be between five and ten seconds for each item, overall, approximately two minutes. The time allocated for the test was twenty minutes for KSU (16 items) and seventeen minutes for UQU (14 items).

IV.5.10.9 Presentation

The tests involved in this research, which were either made by EFL language teachers or derived from real standardised examinations, were neither modified nor proofread. This was so that the two types of tests examined were typical representations, even
though there were some linguistic errors in the teacher-made set. The following were the only exceptions:

(a) In the Cambridge examinations, alternatives were presented next to each other and each was labelled by an emboldened capital letter followed by a space and then the word, as in the below example. In the Michigan examinations, the alternatives were presented underneath each other. Each one was labelled by a small letter, followed by either a dot or a bracket, spaces and then the word. The following is an example (Briggs, Dobson, Rohlck, Spaan and Strom, 1997):

- It’s easy to get a library card in New York; everyone is _________ to apply.
  
  a. mandatory
  
  b. legible
  
  c. eligible
  
  d. inevitable

There were also other diverse modes presented by the teachers. For the sake of consistency, all the alternatives were presented with capital letters followed by dots and then spaces. This was done because if ‘A’ had been presented without some mark following, at first sight it could have been read along with the next word. The alternatives were presented next to (not underneath) each other, as in the Cambridge examinations, to save space.

(b) The Cambridge examinations do not provide examples and the testees are instructed in a way that differs from that of the Michigan examinations where examples are usually given. The teachers also used different instructions. Each method has its own positive and negative points, but for this research a brief teacher’s instruction was used since the subjects were already familiar with this type of format.
IV.5.10.10 Procedure

The subjects took the study tests after receiving written and spoken instructions (details about the instructions are in IV.4.2). Whilst the participants were engaged in the test, the researcher was in the control monitoring room at the back of the laboratory with the technician. From that position, could hear the subjects and talk to them as individuals, groups, or all together at the same time and they could do the same by using a button in their booths. The participants were told when there were five minutes left of the time allowed. They were informed again when only one minute remained for them to finish the test (as done by most test administrators there). When the time was up, the researcher asked the test-takers to stop. By that time, the vast majority of them had already finished. Only two subjects had not completed all the last item, when the researcher asked them to stop the test, but they had attempted more than 92% of the test.

The number of participants who took the test was more than originally considered for the study. It had been planned to use thirty-two participants but the number which actually sat for the two sessions of tests and performed the introspective phase was forty. There were an additional two high proficiency students and two low proficiency ones in each session. The extra subjects were included in case some of the intended subjects did not turn up on the day of the study test or something went wrong with their recording machines. If everything went as planned (as actually happened), it would be known in advance who could be withdrawn as the extra participants were from the candidates who fell in the gap between the high and low proficiency subjects (details about the gap are given in IV.4.6).

It was clear that the participants were dealing with the test and with the situation in the same way as they would typically have done in an actual test. Apart from the provision of introspective reports, which was necessary for the study, the two test-taking sessions were carried out in the same conditions as those with which the subjects were familiar.
IV.6 Difficulties

The researcher being a member of staff at KSU greatly facilitated most of the data collection procedures. A number of obstacles were faced, however, in some stages of the data collection. The following were the most difficult problems encountered.

IV. 6.1 Listing the test lexical items

Obtaining the targeted lexical items of the PMT (as explained in IV.4.10.2) was not a straightforward task. This was because, when the targeted lexical items were isolated they were likely to be understood as meaning something different from what they meant in the original context. Consequently, phrases and collocations in the list were used for those words that were more likely to be interpreted differently from their original context. However, even when some lexical items were written with appropriate collocations or in sentences, other lexical items might still have had different meanings.

In order to avoid any mistakes at that stage of the study, the language instructors read the original test items after they had constructed their tests and before the items were approved for use as part of the study test. This was done to establish whether or not the target lexical items in the PMT represented the meanings understood by the teachers when they identified them in the list. As a result, the following item was discarded from the study test (Local Examinations Syndicate, University of Cambridge, CPE, June,1999: p2):

- In ………… course, the truth will be revealed.

    A certain       B due       C real       D belated

The meaning that the teacher understood from the target word included in the list (i.e., immediately payable) did not match the meaning targeted by the PMT item (i.e., at the appropriate time).
IV. 6.2 Immediate interviews

Ideally, the subjects needed to be interviewed immediately after doing the test, so that their mental activities were still fresh in their minds. However, as each interview was to take approximately forty-five minutes and as there were twenty interviewees in each group, this was not possible. Every effort was made to interview as many subjects as possible immediately after the test. The rest of the interviews were carried out on the following two days.

IV. 6.3 Audio-recording

An unexpected problem emerged when the introspective and retrospective reports were being elicited in a language laboratory at KSU. It was found later that the audiotapes could not be played back by using a normal tape-recorder. The language laboratory had recorded on four tracks of the tape, two for the students’ voices (side A & side B) and two for the master-presenter’s voice. The interviews had, therefore, to be conducted in the laboratory in order to enable both the researcher and the subjects to listen to the audiotaped reports. It was also necessary to transcribe as many audiotapes as possible. The untranscribed ones were left with the technician, who later found a way to transfer them onto other tapes on two tracks and mail them to the researcher in England.

IV.7 Summary of Chapter

This chapter has set out the questions the study attempted to answer. It also reported the lack of adequate data collected by the pilot study, which pointed the way to improving the data gathering methods in the main study. The background of the subjects who took part in the study was explained and it was shown that they shared a similar background and schooling history. Details were given about the eight instruments used to elicit the data required for the study and the process of developing the study test was described. The data was gathered via various VP methods, namely, introspective reports and two other supportive verbal-base retrospective methods. Other instruments were used to determine three independent variables potentially affecting TTS. These variables were:
The thirty-two subjects took two types of tests, PMT and TMT as a repeated measure. Finally, in this chapter, the difficulties the researcher faced in gathering the data were reported.

- Type of test (i.e. PMT and TMT).
- General level of proficiency in English.
- Level of lexical proficiency in English.