Data Analysis

Khalid Addamigh, PhD

King Saud University, Saudi Arabia

This chapter focuses on the process of analysis which the data underwent. It comprises information about the protocols’ transcription, segmentation and the development of the TTS classification and the validity and reliability of the analysis and details about the quantitative analysis. The findings of the analysis, involving the TTS taxonomy scheme, will be presented in the following chapters. Figure 5.1 provides a summary of the five stages of the data analysis process.

Figure 5.1: The main steps through which the think-aloud data were analysed
The following are details about the data analysis steps.

V.1 Transcription

The examinees’ spoken data was transcribed, in order to be dealt with through paper-based analysis. Since every piece of protocol might indicate a certain behaviour related to a strategy use, the data was transcribed verbatim. The following typefaces and markers were used in this step (Chapter VII provides examples):

- Text written in *Monotype Cursiva font* indicates that the original text was in Arabic (in the case of translation). The Times New Roman font was used for what was said in English.
- Where letters are repeated, for example in *too*, it means that the examinee elongated the word.
- Bold typeface indicates that the examinee was stressing the word more than normally.
- One tilde ~ indicates a relatively short pause.
- Two tildes ~~ indicate a longer pause.
- Three tildes ~~~ indicate a relatively long pause.
- Text written in [square brackets] indicates that that information was derived from the supportive instruments, retrospective report or interview.
- Anything written between {curved brackets} indicates that the examinee was interrupted by something.
- Text in <pointed brackets> indicates a comment by the researcher on the test taker’s protocol.
V.2 Translation

The vast majority of the participants used their mother tongue to voice their thoughts. To avoid any problems that may take place in translation, the protocols were dealt with in their original language by the researcher during the process of the analysis. However, it was necessary to translate some of the Arabic protocols into English for the purpose of having other experts, who may not speak Arabic, check the content validity.

V.3 Segmentation

After the data was transcribed, step ‘a’, it had to be examined carefully to infer the strategies used. Every distinct fragment of behaviour that provided evidence of a type of strategy was highlighted. If a certain piece of behaviour in the protocols exhibited more than one type of strategy, it was given two colours and each of the strategies involved was coded. Such cases were, however, limited. Repetitions of the same strategy used more than once for a given item by a person were noted, though in the quantitative analysis a strategy will only be counted once per item. Before the researcher embarked on any individual test protocol, he read the protocol report carefully more than once to obtain an overview of each subject’s test situation and to identify the general strategies used that were not related to any specific test item (eventually labelled as category ‘a’, see Chapter VII for these strategies). The coding of each test item protocol was reviewed to make sure that no strategy was missed or misidentified. The work at this stage was done by the researcher and based on the Arabic version of the protocols. Two experts read samples of the translated (and the Arabic) version and they were the contacted for discussion and their views compared in order to arrive at a proper classification. Moreover, when the researcher found it difficult to classify or identify (or translate) a piece of behaviour, he sought the help of
his supervisor or the expert who helped to check the reliability (more details are in V.7).

V.4 Classification

Stage ‘c’ of the analysis process was undertaken to identify the complete range of strategies of which the data provided evidence. At this stage, after much reviewing, these strategies were classified into a robust TTS taxonomy scheme, as follows.

V.4.1 Developing the taxonomy scheme

The taxonomy scheme was developed in the following ways.

1- First, six protocols of ‘high’ and ‘low’ language proficiency subjects were examined in an effort to initially classify the range of different ‘types’ of TTS used by the test takers. See V.3 for details.

2- The types of strategies found in step ‘1’ were then grouped into seemingly logical categories according to the particular step of the test-taking process that they served. Details on categorisation are in V.4.2.

3- The rest of the protocols were analysed, as explained in V.3, and various new types of strategy were added to the scheme, i.e. the TTS scheme was updated and modified several times, during which each protocol was looked at.

4- Each single distinct strategy was called a ‘strategy type’. The taxonomy scheme, therefore, comprised the ‘types’ of strategies used by the subjects.

5- Each type of TTS included in the taxonomy coding strategy scheme was labelled with a letter and a number. The codes were written alongside the corresponding protocol segment in a column prepared for that purpose when making the transcriptions.
6- With the aim of obtaining as complete and systematic a scheme of TTS as possible, the protocols were rechecked all over again, to ensure that every possible type of strategy that the data provided evidence for was included in the taxonomy.

7- Each occurrence of a strategy type, the repeated use of the same strategy, was called a 'strategy token', using Scholfield’s (1995) term.

8- Finally, the validity and reliability of the scheme were examined. See V.6 and V.7 for details.

We must state explicitly here that it was not always a clear-cut job to classify every segment of the think-aloud thoughts, as some aspects of the test takers’ behaviour could be interpreted in different ways. When the researcher experienced such difficulty, he sought the help of his supervisor or the expert who helped to check reliability. Moreover, the process of the developing the taxonomy scheme, required a long process of development and modification based on painstaking analysis of the protocols.

V.4.2 Categorisation

A general look at previous classifications of strategies shows that strategies have been broadly classified in the following two ways:

- The first approach is to group the strategies that can occur according to the particular skill or processing domain they serve, for example, strategies related to communication (Littlemore, 2001), vocabulary learning (Schmitt, 1997), writing (Young, 2000), pronunciation (Derwing and Rossiter, 2002), reading (Kozminsky and Kozminsky, 2001), and then to sub-classify if necessary.
• The second is to sort strategies according to their ‘kind’ of function in cognitive or affective terms. This approach classifies strategies according to a binary approach, based on broad distinctions thus, we find ‘cognitive’ and ‘metacognitive’ strategies (O’Malley and Chamot, 1990), ‘direct’ and ‘indirect strategies’ (Oxford, 1990), ‘introversion’ and ‘extraversion’ strategies (Wakamoto, 2000).

In this study, the first approach was adopted with the relevant skill processing domain being test-taking strategies, although reading was also relevant, because of the stem provided for each test item. Then, the strategies were subcategorised into seemingly logical groups according to the particular phase of the problem they served to tackle. This was done using some awareness of reading and test-taking strategies suggested by previous research, e.g. Nevo (1989) and Anderson (1991), (details about them are in Chapter II.9).

Eventually, after many exchanges of ideas and discussion with other experts, it appeared that the strategies generally fall into the broad categories outlined in Figure 5.3. Apart from the strategies that seem to be largely independent of specific test items, the test taker might use some strategies to solve problems or limitations that occur in understanding the stem construction, the meaning of the gap word, the alternatives or part of them or using the mental lexicon in LTM. This corresponds to path ‘b’ in Figure 2.4 (see Chapter II, 6.3). If the stem meaning and the gapped meaning were known and the meanings of the alternatives were retrieved so as to be temporarily in STM, the answer picked would be based on matching the required meaning for the gap with the meanings of the alternatives (see II.5.1.3 and II.7 for more details). Hence, ideally,
when an individual attempts an MCFGV test item, one might imagine the test-taker would follow the steps in Figure 5.2:

![Figure 5.2: 'Ideal' steps for attempting an MCFGV meaning test item](image)

Although not many of the subjects overtly inferred the missing lexical meaning, stage 2, or give evidence in their VP of guessing, (before reading the alternatives), what sort of word the gap needed we found that the test-takers did follow a similar pattern to Figure 5.2. This can be seen in Figure 5.3.
General strategies which are not associated with an individual item

Step ‘1’

Reading comprehension of the stem & inferencing what the gap needs

Success  |  Reading C. strategies  |  Lack

Step ‘2’

Examining the alternatives for the gapped word

Success  |  strategies  |  Lack

Step ‘3’

Selecting an answer

The answer is known  |  strategies  |  The answer is unknown

Strategies used after choosing the answer

Figure 5.3: Model of the actual method used by the test takers in tackling MCFGV test items
Based on this tentative model of how test-takers respond in a MCFGV test, the strategies found in the data were grouped into the following categories:

(a) Strategies for managing the test as a whole, which may occur before, during or after the strategies associated with individual items, categories covered in category ‘b’ onwards.

(b) Strategies for stem comprehension.

(c) Strategies for handling the gap.

(d) Strategies for handling the alternatives.

(e) Strategies for choosing the answer (positive and negative lines of attack).

(f) Strategies used after choosing the answer.

Abbreviations were used for the strategies to facilitate the analysis. The codes used to label the strategies were based on the above alphabetical categorisation, associated with a number to indicate each individual strategy. See Chapter VII for a full account.

V.5 Application

After the development of the TTS taxonomy scheme, step ‘d’, the protocols were rechecked all over again to apply this taxonomy to the data. Since the coding scheme underwent a long process of development, the coding of many protocols was rechecked and changed several times. Initially the protocols were read many times for step ‘d’, and later several more times for step ‘e’. Some previous researchers have placed a quantitative restriction on the application of a categorisation, i.e. a strategy was only recognised and coded if it occurred a certain number of times (e.g. Olshavsky, 1976/7). In this study, however, no quantitative criterion was set for coding and counting the strategies. Hence, every strategy was considered, regardless of the frequency with
which it was repeated. This was because one of the aims of this research was to uncover as wide a range of strategies used as possible, in an area little studied before.

V.6 Validity of data and analysis

During steps ‘a’ transcription, ‘b’ translation, ‘c’ segmentation, ‘d’ development, and ‘e’ application, considerable emphasis was placed on increasing the validity of the data and procedure whenever possible.

V.6.1 Validity from the triangulated types of evidence

The circumstances under which the concurrent think-aloud protocols are carried out increase/decrease the validity of the data gathered. Green (1998) states:

> Validity is maximised by adhering to certain principles in the procedure. First, it is important to ensure that appropriate instructions are used in order to guide the production of the verbal report. Individuals must be discouraged from trying to explain or rationalise their thoughts. The verbal reports should ideally be produced as the task is being carried out, with minimum intervention from the researcher (Green, 1998: 11).

These principles, and others, were followed. See IV.4.5.3 for details. However, to further increase the validity of the data gathered by the concurrent think-aloud, the retrospective reports and interviews were used as complementary or supportive evidence for the introspective report, the main instrument. The interpretation of the data elicited by the introspective report was looked at simultaneously with the data elicited by the other two instruments. For more details see Chapter IV.5.1. The feedback from these three instruments, therefore, was neither coded nor analysed independently. This was in order to clarify as much as possible anything that was incomprehensible when the students were reporting concurrently with the test, so as to increase the validity of the obtained data and the result of the study. The following is an example of how the
two follow-up instruments helped to clarify something reported initially without sufficient clues to what strategy might have been involved with respect to the choice of the D alternative.

The test item was:

The killing disease __________ in so many countries.

A. went  B. went ahead  C. spread  D. faced

The introspective report was:

The killing disease space in so many countries. The killing disease ~ in so many countries… disease?! Disease .. I do not know what it means, but I think I have come across it somewhere.. It is familiar ~. aaa ~~~ I think the required word is ‘D’, faced.

The related retrospective report was:

I have chosen ‘D’, guessing.

The interview was:

The researcher: Why did you choose this alternative? The test-taker: since ‘A’ and ‘B’ are synonyms, I thought the answer could not be one of them, as there should not be two correct answers in the alternatives. When I was left with ‘C’ and ‘D’, I chose ‘D’ only because I had not chosen the letter ‘D’ in the preceding items, but the other one could also be correct. The researcher: Do you know the meaning of ‘C’ or ‘D’? The test-taker: No.
V.6.2 Validity of the analysis procedures

The data analysis procedure itself can decrease the validity of the findings. Thus, the following were considered when analysing the data.

1) If the transcription did not reflect exactly what was detected by the audiotaping, it could reduce the validity of the data. Therefore, every possible effort was made to transcribe the spoken data exactly. Markers and typeface, details in V.1, were used in the written protocols in an attempt to reflect closely the content of the spoken protocols.

2) The vast majority of the participants used their mother tongue to voice their thoughts. To avoid any problems that might take place in translation during the development of the analysis, the researcher dealt with the protocols in their original language. This which would increase the validity of step ‘b’ onward. However, it was necessary to translate some of the Arabic protocols into English for the purpose of allowing others to check validity and reliability.

3) The content validity of the strategies classification, step ‘d’, was verified by two experts: (a) the supervisor of this study, Mr. Scholfield, who examined samples of the translated protocols and other protocols that were especially difficult to classify and (b) the expert who checked the interjudge reliability, as he examined samples of the Arabic version of the classification system.
V.7 Reliability of segmentation and application

The vast majority of the studies reviewed do not indicate clearly the procedures used for checking the reliability of protocol coding systems, “or, indeed, do they give explicit information about validity (e.g. Olshavsky, 1976/7; Block, 1986; Nevo, 1989; Anderson, 1991; Upton, 1997; Orr, 2002). In this study, when a final version of the classification had been developed, i.e. when the strategy types had been identified, categories established, examples extracted and a definition given of each strategy, the reliability of the coding system used in the study was checked. The reliability of the coding system was examined via two methods of reliability check: (a) interjudge and (b) intrajudge.

V.7.1 Interjudge reliability

In this reliability check, two coders code certain protocols and a comparison between the two coders is made. Coders are unlikely to be asked to segment and code the verbal protocols. Instead, the coders are more likely to be asked to code protocols once segmentation has been carried out (Green 1998). Hence, an expert, who was considered as the second judge, was asked to perform the reliability check. He was aware of the objectives of the procedure as he had experienced this requirement in his PhD research on L2 word attack strategies. A sample of three transcript protocols (9.38%) was randomly chosen. Two of the chosen protocols were form subjects of a high language proficiency level and the other one from the low proficiency group. The researcher, who was the first judge, coded the three protocols. The protocols given to the second judge, which were in Arabic, were not encoded and so he had to decide what was and what was not evidence of a strategy, as well as which strategy it was. He was given the list of strategies with examples and descriptions elaborated by the researcher and asked to analyse the data and code it. After he had finished the analysis, the two judgments
were compared, to look for agreement and disagreement. Table 5.1 shows the number of strategies identified in each protocol by the two judges.

<table>
<thead>
<tr>
<th>Protocol</th>
<th>Strategies tokens in the 1st judge</th>
<th>Total agreement between the two judges</th>
<th>Differences in coding</th>
<th>Strg. not recognised by 2nd judge</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>108</td>
<td>87</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>80</td>
<td>66</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>95</td>
<td>76</td>
<td>6</td>
<td>13</td>
</tr>
<tr>
<td>Total</td>
<td>283</td>
<td>229</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.1: Interjudge Reliability Coding

From Table 5.1 it can be seen that the first judge, the researcher, found 283 tokens of strategies over the four sample protocols and the second judge reported 229 tokens of strategies. Following Scholfield’s formula (1997a) the interjudge reliability coefficient was calculated as follows:

\[
\text{N\^{} of strategies tokens placed in the same category by all judges} = \frac{(229)}{(283)} \times 100
\]

The result was 80.92% agreement, which showed an acceptably high degree of reliability.

V.7.2 Intrajudge reliability

The present researcher coded two introspective protocols. One of them was of a ‘low’ language proficiency subject and the other of a ‘low’ proficiency one. In the first reliability check there were two protocols from the high proficiency subjects and one from the low proficiency group. After two weeks, without looking at the first coding, the researcher went over the two protocols again and recoded them. The first and
second instances of coding were contrasted. It was found that in the first, 100 strategy
tokens were identified but 94 in the second. The number of tokens that were coded the
same in the first and second coding was 92. Hence, the total agreement between the
first and second coding was 92.

The intrajudge reliability of the codification was calculated according to Scholfield’s
formula (1997a) by dividing the number of strategies coded the same in the first and
second times of coding by the number of strategies coded in the first codification, as
follows:

\[
\frac{N' \text{ of strategies coded the same by researcher in the first and second coding} = 92}{N' \text{ of strategies coded by the researcher the first coding} = 100} \times 100
\]

Accordingly, the intrajudge reliability coefficient for this study was 92% agreement,
which proved the coding system to be highly reliable. Both methods of reliability
check, therefore, supported the reliability of the coding system. Furthermore, the use of
two ways of measuring the reliability further increases confidence in the reliability of
the findings.

V.8 Principles for quantitative analysis
The following points were taken into account when counting strategies for the
quantitative analysis in this study:

1- The strategies of category ‘a’ (located in Chapter VII) were counted as ‘occur’
or ‘not occur’, ‘1’ or ‘0’, for each test-taker. This category concerned the
management of the test as a whole and mostly consisted of metacognitive
strategies. For groups ‘b’ onwards, which were associated with individual test
items, the frequency of each strategy type was based on the number of times (strategy tokens) that each one was repeated for different test items.

2- Repetitions of the same strategy used more than once on the same test item by the same test-taker were counted as one token. The maximum possible frequency of a strategy for any subject was the number of test items involved in a test protocol. For example, if a subject reread the stem of an item repeatedly many times and, during these rereadings, used strategy c.17, skipping the gap, several times for the same gap, then strategy c.17 for this item in this protocol would be counted as one token. This was because some test-takers used a particular strategy type repeatedly in their protocols. This would give an impression in the quantitative result that this strategy type was commonly used by the test-takers, whereas in fact only a few subjects used it, but used it constantly.

3- When the researcher identified in the introspection reports an application of a strategy for a certain item and then this use was evidenced again for the same point in the retrospective and/or the interview phases, this use was calculated as one frequency token.

4- The numbers of test items taken by the two groups of test-takers were not the same in the two tests: there were 16 test items in the KSU test and 14 in the UQU test. To ensure comparability, the strategy frequencies for strategy types ‘b’ onwards were all converted to percentages of the items in the tests. For example, if strategy e54, selecting the odd alternative in form, was used by someone twice in a test of 16 items, it was recorded as used in 12.5% of the
items, whereas if it was used in 2 items out of 14, it would considered on being used in 14.3% of the test items. Another example is, if a test-taker used, for the gap strategy c20, a filler word 14 times in a test of 14 item, it would mean that the strategy was used in 100% of the test items, whereas if the test-taker used it 14 times in a test of 16 items, the strategy would be considered to be used in only 87.5% of the test items.

5- Since the two groups of subjects sat two tests of different numbers of items, all the correct scores of the subjects were converted into a percentage of the relevant maximum possible score. This was, again, to ensure fair comparison of the test scores for each subject. For example, if the items are equally weighted, a score of 14 on a test of 14 items (100%) is in effect higher than a score of 15 on a 16-item test (93.75%).

V.9 Summary of chapter

This chapter introduced the methods used in analysing the main data. It gave details of how the examinees’ spoken data was transcribed in order to be dealt with by means of paper-based analysis. Information about the need to translate some of the protocols followed. After transcribing the testees’ spoken language, the data was checked carefully to infer the strategies used in the test. Every distinct fragment of behaviour that provided evidence of a type of strategy was marked. This stage was undertaken to identify all the strategies used in the data. The following stage classified and categorised these strategies into a TTS taxonomy scheme. This required a long process of development and modification based on a painstaking analysis of the protocols. The chapter highlighted the fact that it was not always a clear-cut job to classify every segment of the think-aloud thoughts, as some portions of the test-takers’ behaviour
could be interpreted in different ways. It was pointed out that when the researcher
experienced such difficulty, he sought the help of his supervisor or the expert who
helped to check the reliability. Details were then given about the methods used in
previous studies to classify strategies. This study followed the approach that first
groups the range of strategies that can occur according to the particular skill or
processing domain they serve, TTS in our case, then subcategorises the strategies into
seemingly logical groups according to the particular problems they were used solve.
These were: (a) strategies for managing the test as a whole, (b) strategies for stem
comprehension, (c) strategies for handling the gap, (d) strategies for handling the
alternatives, (e) strategies for choosing the answer and (f) strategies used after choosing
the answer. After the development of the TTS taxonomy scheme, the scheme was
applied to all the protocols. The chapter also pointed out some issues regarding the
validity of the findings. The methods used to check the reliability of the coding
taxonomy scheme, which indicated a high degree of reliability, were also discussed.
Finally, information about the quantitative analysis used in this study was presented.