CHAPTER FOUR

RESULTS OF THE STUDY

This chapter presents the results of the data, collected and analyzed as described in the methodology chapter. The results of this study are organized according to the data analysis sequence in order to answer the research question and discuss the salient themes of the present study. Therefore, the first section will present the listening test scores and participants' understanding of listening texts, and provide explanation for these results based on data of retrospections and interviews. This will be followed by the analysis which focuses on finding out the relationship between test scores and the corresponding scores of understanding for the listening tests which utilized MC and SQ test formats. Finally, data will be reanalyzed in detail to find out the underlying reasons which have caused these results.

Results of Scoring

Scoring in present study referred to scoring of tests and scoring of listening text understanding. The possible test scores ranged from 0 to 10, and possible scores of listening text understanding ranged from 0 to 20 (each of the total of 10 items gives a score of 0, 1 or 2). The results will be presented in the following

subsections.

Listening Test Scores

Participants in Listening Test 1 (MC) obtained higher test mean scores than those in Listening Test 2 (SQ) (see Table 4).

Table 4 Participants' test scores

Listening Test 1 (MC)		Listening Test 2 (SQ)	
Participant	Test Score	Participant	Test Score
1	6	11	5
2	3	12	3
3	3	13	6
4	4	14	3
5	2	15	3
6	5	16	3
7	5	17	2
8	4	18	5
9	7	19	4
10	6	20	2
Total	45	Total	36
Mean	4.5	Mean	3.6

Understanding of the Listening Texts

Participants in Listening Test 1 (MC) achieved higher mean scores of understanding than participants in Listening Test 2 (SQ) although the two tests used

the same listening texts (see Table 5).

Table 5 Participants' scores of understanding

Listening Test 1 (MC)		Listening Test 2 (SQ)	
Participant	Score of Understanding	Participant	Score of Understanding
1	10	11	6
2	7	12	4
3	4	13	8
4	6	14	3
5	2	15	3
6	7	16	5
7	2	17	2
8	5	18	5
9	6	19	5
10	6	20	2
Total	55	Total	43
Mean	5.5	Mean	4.3

Comment on Differences in Scores

The two mean scores in Listening Test 1 (MC) which came from answering test questions and understanding of listening texts were higher than those in Listening Test 2 (SQ), which naturally raises questions about the causes of this phenomenon.

There were two possible answers for these questions. The differences in scores could be possibly caused by the different average listening proficiency of the two groups of participants. Participants in Listening Test 1 (MC) could possibly have higher listening proficiency than those in Listening Test 2 (SQ) although they

have been selected and arranged into these two tests according to their NMET scores (see Table 1). Another possible answer was that this phenomenon was due to the influence of method effects of test formats. In addition, there was evidence in the retrospective verbal report and interviews that this result was actually caused by the method effects of MC and SQ test formats.

To illustrate, in her retrospective report, participant 13 who took part in Listening Test 2 (SQ) said she could have performed better on the same test item with MC format.

> This man, this man, er, said, the other one, like, er, to do something. Then this woman, this woman said: "according to his opinion, it's not this; it should be a different way." So the meaning of this woman should be, disagreed with the man's opinion. Er, although I didn't hear the answer of this question clearly, if I was provided with a MC question, I should be able to choose the correct option. Anyway, the woman's opinion should be opposite to the man's opinion. They were different opinions, because she refuted the man. Er, that's all.[1]
>
> (Participant 13, SQ, Retrospection, Item 7)

Similar opinions were reported in interviews by the participants who took Listening Test 2 (SQ). They emphasized that if they were in the test with MC format, they could obtain better scores.

[1] Participants have produced various grammatical mistakes, wrong choices of words, incoherent sentences in their retrospective verbal reports although the language they used to report was their native language (Chinese).

Participant: er, teacher, er, I, I only got 5 in the test, but I believe, er, if, if I was in that, that test with MC format, at least, er, I could get 7 because I could choose correct options for question 6 and question 10.

Researcher: Why do you believe you can choose the correct option?

Participant: This, this test was our first test in the college, so we talked about it in the dorms, and we checked answers with each other. Er, so I found I could choose right option in another test, er, because I heard some key words, guessed, and if I read the options, er, compared with the options, I would recognize the correct options.

Researcher: Could you tell me which test format you prefer?

Participant: Multiple-choice questions. Multiple-choice questions are, are, er, easier. I can predict what I'm going to hear according to options. Er, so I feel it's easier to understand. In addition, er, I can eliminate some options too. Whether or no, ah, er, I, I have twenty-five percent of chance to be correct by guessing. (Participant 11, SQ, Interview)

According to the interviews, participants believed MC format could decrease the difficulty of listening tests, because reading the options could help predict what they would hear next, increase their understanding of listening texts, and help them catch key words for answering questions. After listening, they could analyze options and eliminate the possible incorrect choices according to what they heard and their background knowledge. All of them indicated the important role of guessing.

However, while reading their scored test papers at the beginning of interviews, some participants in Listening Test 1 (MC) soliloquized quietly about questions which they thought correct but wrong, and which they thought wrong but correct. When they were asked, they claimed that they felt they had understood the conversations but their answers turned to be wrong.

Relationship between Tests Scores and Understanding Scores

The purpose of present study was to investigate which test format, MC or SQ, best measured listening comprehension. In other words, the accuracy of MC and SQ as measures of test takers' listening comprehension should be examined. In order to compare MC and SQ formats, the question of whether test takers' performances (test scores) on listening tests were consistent with their understanding of listening texts (understanding scores) needed to be examined.

To examine which test format best measures a trait is actually investigating the reliability and validity of test formats. Alderson et al. (1995, p. 294) explained that "reliability is the extent to which test scores are consistent: …, the reliability of subjective tests is measured by calculating the reliability of the marking". These researchers also pointed out that MC test format had higher reliability than SQ test format (1995, p. 88). However, due to the highly reliable marking processes for SQ questions in the present study, the difference of reliability on test format between Listening Test 1 (MC) and Listening Test 2 (SQ) was minimized to a very low level.

Validity is defined as "the appropriateness of a given test or any of its component parts as a measure of what it is purported to measure" (Henning, 1987. p. 89). Alderson et al. (1995), Bachman (1990), Henning (1987), and Hughes (2003) all emphasize the importance of validity because it is the extent to which tests measure what they are supposed to measure. In other words, validity is central to testing. In terms of the definition of validity, the validity of MC and SQ test formats in this study should be the accuracy of measuring test takers' listening comprehension. In order to examine the accuracy of MC and SQ formats on measuring listening comprehension, the relationship between tests scores and understanding of texts should be analyzed.

Bachman (1990), Bailey (1998), Henning (1987), and Hughes (2003) suggest the use of correlation coefficient to analyze test data for examining test validity. Bachman (1990, p. 259) explains correlation as "a functional relationship between two measures". He also states "to say that two sets of scores are correlated with each other is simply to say that they tend to vary in the same way with respect to each other" (1990, p. 259). Through the analysis of correlation, the researcher can find out whether test takers' performance is related to their understanding of listening texts. In other words, the purpose of correlation analysis is to identify whether participants' test scores were high when their scores of text understanding were high, or whether these two sets of scores were inversely related, or whether there was no relationship between these two sets of scores.

Results of Correlation Analysis

Correlation Result of Test 1

The test scores obtained in Listening Test 1(MC) were compared with listening text understanding scores by utilizing SPSS software package with Bivariate Correlation method (using the Listening Test 1 scores, from Tables 4 and 5). The scattergram of the two sets of scores is shown in Figure 2.

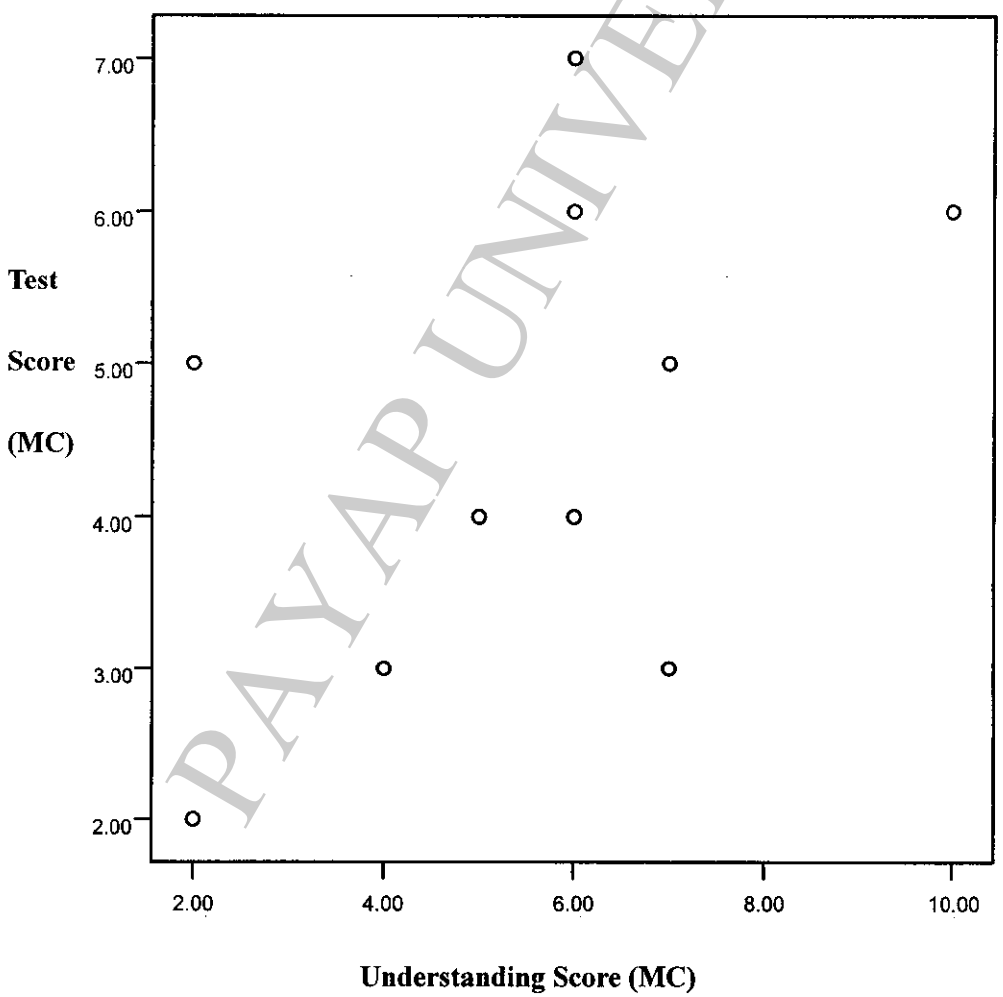

Figure 2: Scattergram of test scores and understanding scores in Test 1 (MC)

From the scattergram, visually the two sets of scores appear only weakly related. This could suggest that the correlation between the two sets of scores is not strong. The data was further analyzed by Pearson's correlation coefficient, with the result shown in Table 6.

Table 6 Correlation of scores in Listening Test 1 (MC)

	Understanding Scores of Listening Test 1 (MC)	
Test Scores of Listening Test 1 (MC)	Pearson Correlation	0.480
	Sig. (2-tailed)	0.160
	Sum of Squares and Cross-products	16.500
	Covariance	1.833
	N	10

The result showed that Pearson's *r* was 0.480, which could be interpreted as a moderate correlation. As Bailey (1998, p. 113) say, the closer the value of *r* is to the whole number 1.00, the stronger the relationship between the two variables. Therefore, correlation coefficient *r* values of 0.85 to 0.99 would be considered "strong", 0.70 to 0.84 "moderately strong", 0.45 to 0.69 ranges are "moderate", and those below 0.45 could be considered "weak to moderate" correlations. According to the description, the relationship between participants' test scores of Listening Test 1 (MC) and text understanding scores of Listening Test 1 (MC) is only moderate. The result suggested that a correlation existed between test scores and understanding

scores for Listening Test 1 (MC) but this relationship was not significant.

Correlation Result of Test 2

The test scores and understanding scores of Listening Test 2 (SQ) were analyzed as before. Figure 3 shows the resulting scattergram.

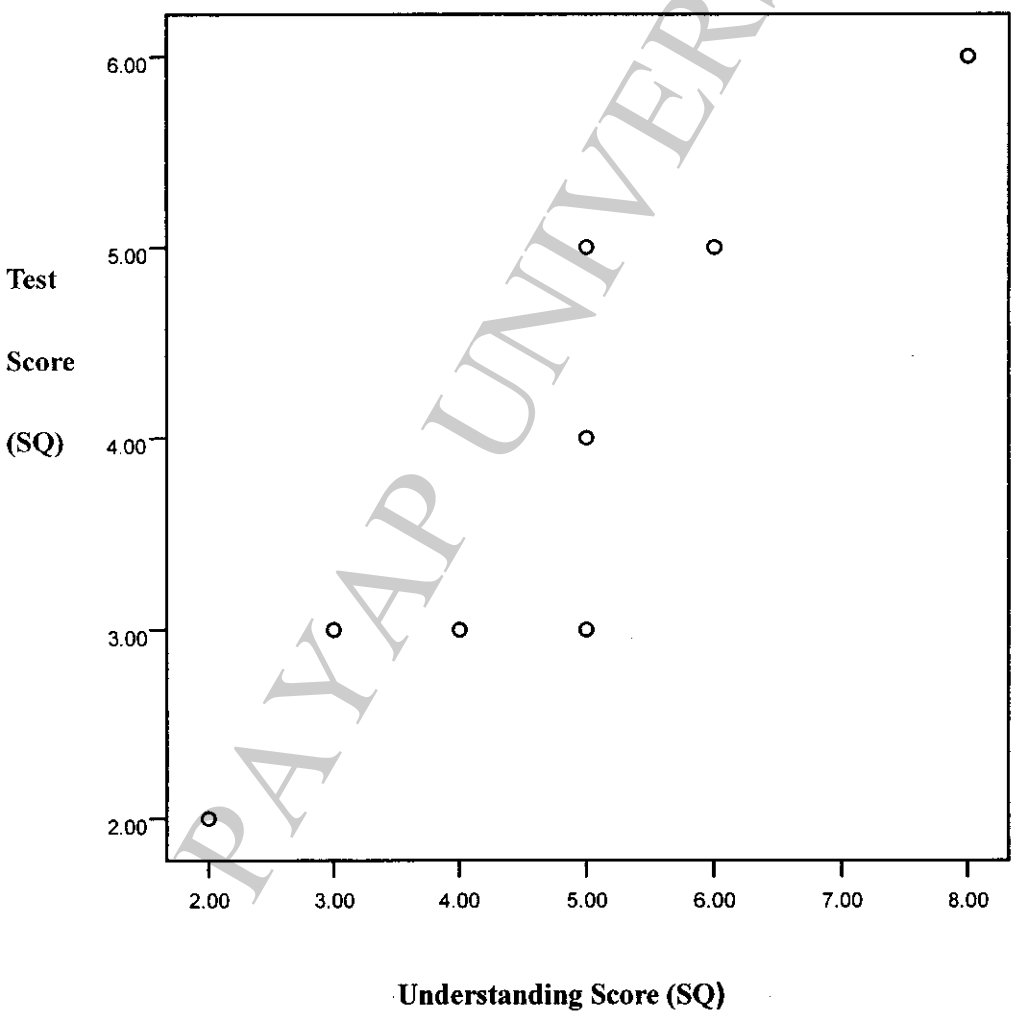

Figure 3: Scattergram of test scores and understanding scores in Test 2 (SQ)

Since the visual appearance indicates that the two sets of scores rise from lowest to highest in the scattergram, the correlation between the two sets of scores is probably strong. From the analysis of data by Pearson's correlation coefficient, the results suggested a strong correlation between the test scores and understanding scores of Listening Test 2 (SQ) (see Table 7).

Table 7 Correlation of scores in Listening Test 2 (SQ)

	Understanding Scores of Listening Test 2 (SQ)	
Test Scores of Listening Test 2 (SQ)	Pearson Correlation	0.924
	Sig. (2-tailed)	0.000
	Sum of Squares and Cross-products	21.200
	Covariance	2.356
	N	10

Note. Correlation is significant at the 0.01 level (2-tailed).

The table shows that Pearson's *r* was 0.924, which can be interpreted as very strong and significant correlation at the 0.01 level (2-tailed). Therefore, the relationship between participants' test scores and understanding scores of Listening Test 2 (SQ) was significant.

Interpretation of Correlation Results

Since Listening Test 2 (SQ) had stronger correlation between its test scores and understanding scores than Listening Test 1(MC), and the major variable was test format in this study, the suggestion is that the validity of SQ test format could be higher than that of MC test format.

However, these statistical results do not provide a complete and satisfactory explanation for the causes of this phenomenon. What caused the difference in correlations between tests performance and listening comprehension between Listening Test 1 and Listening Test 2? Clues to the answers can be found in participants' retrospections and interviews.

Participants reported mismatches between understanding of listening texts and test scores. Based on analysis of retrospective data, in Listening Test 1 (MC), there were six participants who chose wrong options for some test items when they had basically or thoroughly understood the dialogues, and seven participants made the right choices when they did not understand those dialogues. The number of such mismatches in Listening Test 2 (SQ) was much lower than that in Listening Test 1 (MC). For Listening Test 2 (SQ), only two participants failed to score on items when they understood the dialogues and two participants were found to have little or no understanding of the dialogues but wrote down correct answers.

Retrospective reports about the test-taking processes which finally resulted in scores which do not match understanding was produced by participants. For

example, participant 9 and participant 4 in Listening Test 1 (MC) reported their different experiences of dealing with test item 2. The tapescript and question of test item 2 were the follows:

Test Item 2:

Dialogue:
Woman: Look, it says they want a junior sales manager, and it seems like it's a big company. That'll be good. For you might have to travel a lot.
Man: Do they say anything about experience?
Question: What are they talking about?
Options:
A. A job opportunity. B. A position as general manager.
C. A big travel agency. D. An inexperienced salesman.

(Quoted from CET4, the National College English Testing Committee, 2002)

Participant 9 reported how he dealt with the options and information he heard from this item:

This question because of, a moment ago, a moment ago, I was thinking about last question, I didn't hear it clearly. According to what I heard later, content I heard, I heard, er, up, opportunities, so, then I didn't know, so I chose A, a job opportunities. (Participant 9, MC, Retrospection, Item 2)

Although participant 9 had no understanding of this dialogue, he chose the right option and obtained the score. On the other hand, with the same question and same test format, participant 4 was found to choose the incorrect option when she

understood the dialogue:

> This woman said 'look, there is, there is, is, is a company which is inviting applications for a position of manager. Then, I want to have a try'. Then the man asked her if they needed experienced people. The woman said 'didn't know, I didn't talk with them about these'. So of this question, the question of the second dialogue is what they are talking about. According to the above conversation, the answer is a general manager's position. So chose B. (Participant 4, MC, Retrospection, Item 2)

In order to find out more elaborate explanations for these phenomena, each participant's test score on each test item was compared with his or her corresponding understanding of listening text. The results which have been categorized and analyzed will be presented in next section.

Method Effects of MC and SQ Test Formats

In order to compare participants' test scores with their corresponding understanding of listening texts in retrospective report, the data was categorized as follows: test scores obtained only when participants understood the listening texts, test scores obtained when participants had little or no understanding, and test items answered incorrectly despite basic or thorough understanding.

In order to find out whether MC and SQ test formats have different method effects on listening comprehension, the categorized data were analyzed with SPSS software package using the Independent-Samples T Test method. The results showed whether significant differences existed in the categorizations between MC and SQ test formats.

Test Scores Obtained Only When Participants Understood Listening Texts

This categorization referred to those test scores which were obtained by participants when they answered the questions correctly and had basically or thoroughly understood what they had heard. In other words, these test scores were those which test designers hope to get from listening tests: test scores coming from listening comprehension (see Table 8).

Table 8 Test scores coming from understanding

	STTIAU	TIBU	TITU
Listening Test 1 (MC)	32	17	15
Listening Test 2 (SQ)	34	28	6

Note. STTIAU = scoring total of test items answered by participants with basic or thorough understanding; TIBU = test items with basic understanding when correct answers were produced by participants; TITU = test items with thorough understanding when correct answers were produced by participants.

No significant difference exists between the scoring total of test items answered by participants with basic or thorough understanding respectively of Listening Test 1 (MC) and Listening Test 2 (SQ) in this table according to the result of Independent-Samples T Test (t = -0.325, df = 18, p = 0.749 > 0.05, two tailed).

However, based on the numbers of test items with basic or thorough understanding when correct answers were produced by participants respectively, participants in Listening Test 1 (MC) seemed to understand the same listening texts better than participants in Listening Test 2 (SQ). According to table 8, participants in Listening Test 1 (MC) understood more test items thoroughly than participants in Listening Test 2 (SQ). For Listening Test 2 (SQ), participants understood the test items at basic understanding level more frequently.

Comparing the numbers of test items in answered correctly with basic understanding and the numbers of test items answered correctly with thorough understanding (see Table 8 for the two formats) by Pearson's X^2, a significant difference was found (X^2 = 6.491, p = 0.0108 < 0.05). This result indicates that the MC format was associated with correct response to the more test items with thorough understanding than the SQ format, and could indicate that MC format can promote test takers' understanding of listening texts. Combining this statistical result to participants' retrospective reports and interview data, the effect of MC format on prompting test takers' understanding of listening texts was indeed revealed. MC test format was found to be helpful for improving test takers' comprehension by providing

more information on listening texts, giving context of listening, and narrowing down predictions about the coming listening texts.

Many participants reported that they utilized the information which was provided by MC options to predict what they were going to hear, and to combine with the information which they heard in dialogues so that they could understand the content of dialogues better. Participants 1, 3, 8, and 9 reported a similar process of dealing with information from options and information from listening texts. For example, participant 9 said in interview:

> Er, in fact, teacher, many of my answers were guessed. I, when I got the test paper, I read the questions and options first. Because, because, ah, my teacher of English in senior high school taught us to, er, to read and predict what we would hear. Er, after that, er, when I listened to the dialogues, the questions I learnt from options would remind to what I should focus. Then, grasped the key words, and combined the options, chose one from the four options A, B, C, D. (Participant 9, MC, Interview)

When these participants were asked whether MC format could help improve their understanding, they all gave positive answers.

Test Scores Obtained When Participants Had Little or No Understanding

The second categorization was the test scores which participants obtained when they had no or only little understanding of the corresponding listening texts. The result is shown in Table 9.

Table 9 Test scores obtained when participants had little or no understanding

Listening Test 1 (MC)	13
Listening Test 2 (SQ)	2

According to the result of Independent-Samples T Test, a significant difference is found between the two sets of scores which have been obtained by participants when they had little or no understanding of texts. The mean of Test 1 was significantly higher than the mean of Test 2 ($t = 2.819$, $df = 18$, $p = 0.011 < 0.05$, two tailed). This result can be interpreted to be the influence of method effects of test formats. Based on the result of the statistical analysis, MC format was found to have a higher method effect on the test trait because it led to significantly higher test scores which were not associated with understanding of listening texts. Therefore, these scores had no relationship to listening comprehension.

In order to explain exactly which feature of MC test format resulted in this aspect of method effect, the reasons for these scores were analyzed and categorized based on the data from retrospections and interviews.

There were four categorizations of reasons for gaining scores with little or no understanding of texts in Listening Test 1 (MC): making use of general knowledge to guess correct answers, matching words which have been heard in listening texts with options, eliminating impossible options and choosing from the remainder, and totally guessing without any help from listening texts or general knowledge.

Participant 7 adopted the strategy of making use of general knowledge to guess correct answers when she was dealing with test item 6:

Test item 6:

Dialogue:
Woman: Mark's playing computer games.
Man: Should he do that when the final exam is drawing near?
Question: What does the man think Mark should do?
Options:
A. Go on with the game B. Draw pictures on the computer
C. Review his lessons D. Have a good rest

(Quoted from CET4, the National College English Testing Committee, 2002)

Retrospective report: This woman said Mike was playing that electronic, electronic games again. But that man also said, er, he should, he should, er… didn't hear the latter part clearly. But I thought the answer might be C, because we, we usually say don't play games, just study hard. (Participant 7, MC, Retrospection, Item 6)

On the other hand, many participants tried to match the words they caught in

listening texts with words in options. Participant 9 described her test taking process of test item 9 as follows:

Test item 9:

Dialogue:
Man: Excuse me. I'd like to place an advertisement for a used car in the Sunday edition of your paper.
Woman: Ok. But you have to run your advertisement all week. We can't quote a rate for just Sunday.
Question: Where is the conversation most probably taking place?
Options:
A. At a newsstand B. At a car dealer's
C. At a publishing house D. At a newspaper office

(Quoted from CET4, the National College English Testing Committee, 2002)

Retrospective report: This question was a woman, a man and a woman, woman asked about a place. Then it was, er, this woman told him, this woman told him it was, er, …, this woman talked with him. Er, because, because I didn't hear this dialogue clearly, so I guessed it should be, according to its context, and heard in it, the words I have heard, car and, er, and newspaper, and Sunday, I chose newspaper office. B, at, at the newspaper office.

(Participant 9, MC, Retrospection, Item 9)

The technique of elimination was a widely used test-taking strategy in Listening Test 1 (MC). Participants adopted this technique to increase the chances of successful guessing. Participant 7 was a typical example in dealing with test item

9 (see the dialogue and question of test item 9 in the above example, p. 67):

> This, answer B of this test item was at a car, car dealer's. But I, that, in this short passage, in this, er, it was about newspaper, so I could eliminate answer B. Answer C was at a publishing house, publishing house shouldn't be, publishing house should be, how to say, that, publish that newspaper. How could publishing house publish the newspaper? I thought, thought, anyway, answer C was incorrect. Answer A was at a newsstand, and doing things such as talking outside of a newsstand might be more informal. Answer D was at a newspaper office. Just from answer A or D. I thought both, I didn't know which one I should choose from the two answers. It, er, then I chose, I chose answer D. (Participant 7, MC, Retrospection, Item 9)

The last choice of test taking strategies of participants in Listening Test 1 (MC) was blind guessing which means simply guessing without any help from listening texts and general knowledge. Participant 4 described what she has done for choosing options in test item 10:

> Test item 10:
>
> Dialogue:
> Man: I spent so much time polishing my letter of application.
> Woman: It's worthwhile to make the effort. You know just how important it is to get a good impression.
> Question: What do we know about the man?

Options:
A. He wants to get a new position
B. He is asking the woman for help
C. He has left the woman a good impression
D. He enjoys letter writing

(Quoted from CET4, the National College English Testing Committee, 2002)

Retrospective report: The tenth question, er, I didn't hear it clearly, so I guessed answer C. Then I changed into answer A. I thought A probably more important. (Participant 4, MC, Retrospection, Item 10)

Participants in Listening Test 2 (SQ) also achieved two scores by adopting test taking strategies of blind guessing and making use of general knowledge to guess answers. However, the scores they obtain in this way amounted to only two scores. Two participants were successful by utilizing blind guessing and making use of general knowledge to guess answers. Participant 11 reported that the answer of question 4 came from blind guessing:

Test item 4:

Dialogue:
Woman: But what'll happen if it rains? What are we going to do then?
Man: We'll have to count on good weather. But if it does rain, the whole thing will have to be cancelled.
Question: What does the man suggest doing if it rains?

(Quoted from CET4, the National College English Testing Committee, 2002)

Retrospective report: I didn't understand and didn't catch what I heard in

fourth question, listened, couldn't distinguish exactly what he said what he went to do at the end. The only thing I could do was guessing. (Participant 11, SQ, Retrospection, Item 4)

However, participant 11 was quite fortunate because then she wrote down "they would not go" on the answer sheet and was scored to be correct. Participant 18 in Listening Test 2 (SQ) reported similar thinking process of using general knowledge to guess correct answers on test item 6 as participant 7 in Listening Test 1 (MC) as mentioned before (see p. 66). Both of them had the same thinking traces which treated playing computer games as behavior which could be harmful for study and should be stopped. Therefore, participant 7 chose option "review his lesson", and participant 18 wrote down "do not indulge himself with computer games anymore" on the answer sheet. Both of these answers were correct under the context.

Test Items Answered Incorrectly Despite Understanding

In contrast to the last section, this section discusses about test scores which participants should have obtained because they have understood corresponding listening texts. The results of this categorization are shown in Table 10.

Table 10 Test item answered incorrectly despite understanding

	Item Answered Incorrectly	Scores of Basic Understanding	Scores of Thorough understanding
Listening Test 1 (MC)	8	8	0
Listening Test 2 (SQ)	2	1	2

Note. Scores of basic or thorough understanding refer to the understanding levels which were obtained by participants when they answered the questions incorrectly.

According to the analysis with Independent-Samples T Test, a significant difference was found between the two sets of numbers of test items which had been answered incorrectly by participants when they had basic or thorough understanding of texts. The mean of Listening Test 1 (MC) was significantly higher than the mean of Listening Test 2 (SQ) ($t = 2.121$, $df = 18$, $p = 0.048 < 0.05$, two tailed). This result also indicates that MC test format had a greater method effect through misleading test takers to make decisions which were contrary to their understanding.

Data retrieved from retrospections showed that the misled choices in Listening Test 1 (MC) mainly originated from two test items: item 2 and item 5. The main problem with item 2 was the influence of distractors (see Appendices B and C or the above example, p. 60). Option B was quite confusing for test takers especially when they had basic understanding of texts and matched the words they heard from the dialogue with the options. Both of the participants chose option B and reported that they heard "manager's position" in the dialogue.

Similar to test item 2, all the participants who made mistakes on test item 5 decided on option D as the answer.

Test item 5:

Dialogue:
Woman: You took an optional course this semester, didn't you? How is it going?
Man: Terrible. It seems like the more the professor talks, the less I understand.
Question: How does the man feel about the course?
Options:
A. He wishes to have more courses like it
B. He finds it hard to follow the teacher
C. He wishes the teacher would talk more
D. He doesn't like the teacher's accent

(Quoted from CET4, the National College English Testing Committee, 2002)

However, different from test item 2, mistakes on test item 5 were caused by making use of their general background knowledge. All of them understood that the man disliked the professor's classes and could not follow the classes, but all of them finally chose the professor's accent as the reason for the man to dislike the classes. In interviews, participant 3 reported the follows:

> My reason for choosing accent as the answer was because only two options were possible answers. The man said terrible, er, we usually don't, er, we don't say we don't like the course because it is too hard. We usually don't like some classes for special reasons. And, er, er, ah, one of my teachers of English had severe accent problems. (Participant 3, MC, Interview, Item 5)

When she was asked why she did not think difficulty was the reason for students to dislike a course, she answered she would not be a good student if she disliked difficult courses and could not enter into the college. According to her descriptions, most courses in senior high school became very difficult in the third year because students had to deal with endless difficult test questions and assignments in order to be prepared for NCEE.

Participant 11 who was in Listening Test 2 (SQ) showed a possible problem of SQ test format. She has thoroughly understood the dialogue of test item 6, however, she wrote down an answer which was exactly opposite to the correct answer in meaning (see Appendices B and C or the above example, p. 66). In her retrospection, she said:

> In question 6, at the beginning, the woman directly showed her criticism for Mark's playing computer games. Then the man said again, because he would take examinations soon, …[2]. He should continue playing computer games. (Participant 11, SQ, Retrospection, Item 6)

This report showed that participant 11 has understood the dialogue thoroughly. After she said: "because he would take examinations soon", instead of continuing talking about what the man has said, she had a long time pause which might be caused by her thinking or hesitation. Therefore, the following sentence "he should

[2] Participant 11 had a long time pause between the word "soon" and "he" in the retrospective verbal report.

continue playing computer games" was actually talking about her answer to the question "what does the man think Mark should do" since she had written it on the answer sheet. But participant 11 said in interview that she did not know why she answered this question in this way. She said probably because there were no options to narrow down the range of her answer and she just picked up what appeared in her thoughts. This phenomenon might reveal a potential problem of SQ test format: when there is more than one possible answer for a question or the question is vague, test takers could produce a seeming incorrect answer inspired by the understanding of listening texts. In other words, it could suggest that SQ format be only accurate for measuring clearly stated and specific information.

Summary of the Chapter

This chapter presented and discussed the results of this study. The first section presented the results of scoring. The second section discussed the analysis about the relationship between tests scores and understanding scores which was inspired by the analysis in the first section. The findings showed that Listening Test 2 (SQ) gave a higher correlation between its test scores and text understanding scores which suggests that SQ test format had a greater validity for assessing listening. The last section focused on analysis of the method effects of MC test format and SQ test format in the listening tests. The data were categorized into three groups: test scores obtained when participants understood, test scores obtained when participants had

little or no understanding, and test scores not obtained despite basic or thorough understanding. The findings were reached through analysis of statistical data, retrospective verbal data, and interviews. MC test format was found to be helpful for improving test takers' understanding of listening texts. In addition, it was also found to have greater method effects which could result in the imprecise scores for measuring listening comprehension. While on the one hand, MC test format could encourage test takers to guess the answers which produced test scores when there was no or only little understanding, on the other hand, it could mislead test takers to choose distractors when they had understood listening texts. Although SQ test format was also found to have problems, the incidence of these problems was much lower than that of the MC test format. However, one problem of SQ test format was found. When the answer of a question is not a specific point, or the question itself is vague, test takers could produce an incorrect answer despite understanding the listening texts with SQ format.

The next chapter will present the discussion and conclusions of this study. It will also provide suggestions for use of test formats and retrospective reports, limitations of this study and suggestions for future study.