CHAPTER IV
RESULTS AND FINDINGS

4.1 Introduction

This chapter presents the findings from the study which answer the research questions:

1. Can test authenticity and test delivery mediums have any effects on test takers' English reading proficiency and what are their effect sizes?

2. What are the test takers' attitudes towards test authenticity and test delivery mediums?

4.2 The results and findings

The results and findings of this study are reported as follows:

Part I reports the descriptive statistics (arithmetic means and standard deviations) and Analysis of Variance (ANOVA) which ensured midterm and final scores for the 4 groups from the English I course were not significantly different.

Part II reports the testing of ANOVA assumptions: normal distribution and equality of variance

Part III reports the results of the study with regard to whether test authenticity and test delivery mediums have any effects on test takers' English reading proficiency and their effect sizes.

Part IV reports the qualitative aspects of the test takers' attitudes towards test authenticity and test delivery mediums.

Part I

Population

The population of this study was first year undergraduate students at King Mongkut's Institute of Technology North Bangkok during the academic year 2007. There were approximately 1,500 first year students from both Bangkok Campus and Prachinburi Campus. These students were required to take at least 6 credits in English from the English language syllabus to fulfill their Bachelor's degree requirements. The courses required were English I and English II.

Subject

This experimental research employed a 2*2 factorial design. The research design used in this study was the Randomized Block Design. This design was considered appropriate to this study since it was constructed to reduce variance in the data. The sample was, therefore, divided into relatively homogeneous subgroups or blocks (McLinden, D. J. & Trochim, 1998).

Students who took the English I (810301) course in the first academic semester of 2007 were randomly selected and separated into relatively homogenous subgroups or blocks (mentioned on Page 84-85, Chapter 3) according to their English I final and mid-term test grades which were translated from ability scores by: the English Foundation course committee at the Department of Languages, Faculty of Applied Arts, King Mongkut's Institute of Technology North Bangkok. Three groups were formed with grades B-A, C-C+, and D-D+. Normally, the scores obtained from the course were distributed in the shape of a normal distribution with certain statistical characteristics that were known and constant (Bachman, 1990: 72-73). Moreover, the English I test was based on a certain test construct that shared some characteristics of standardized tests. The administration of the test did not vary from one administration of the test to the next. Thus the scores derived from English I midterm and final tests were constant and did not vary and were appropriate for use as students' English ability.

All the randomly selected students were from the faculties of Engineering, Applied Science, Technical Education, Industrial Technology and Management and Agro-Industry.

The number of subjects was 300. There were 75 in each treatment group equally. A one-way ANOVA was conducted to test the significant difference of the subjects'

English language ability. Mean and standard deviation of English language ability of the four treatment groups are shown in Table 4.1.

Table 4.1: Mean and Standard Deviation of English Language Ability of Group One, Two, Three and Four.

	Minimum	Maximum	Mean	Std. Deviation
Group 1	30.18	78.30	52.77	13.08
Group 2	28.00	73.57	51.36	13.47
Group 3	27.14	78.12	51.87	14.21
Group 4	24.81	70.67	50.00	12.18

N = 75 for each group

Table 4.1 shows the mean scores from English I course of the four experimental groups. The mean scores of Group One, Two, Three and Four are 52.77, 51.36, 51.87 and 50.00 respectively. The mean score of Group Four is slightly lower than the other three groups while the mean score of Group One is slightly higher than that of the others.

To test the significant difference between the mean scores of the four experimental groups, one-way ANOVA was conducted at the 95 per cent of confidence level as can be seen in Table 4.2.

Table 4.2: One-way ANOVA of English Language Ability of the Four Experimental Groups

	Sum of Squares	**df**	**Mean Square**	**F**	**Sig.**
Between Groups	300.864	3	100.288	.571	.635
Within Groups	51998.307	296	175.670		
Total	52299.171	299			

$P \leq 0.05$ n = 75 for each group

Table 4.2 shows the ANOVA value of mean scores of the four groups (F=.571), yielding no significant difference at the 0.05 level. This reveals that the English language ability of the four experimental groups is not statistically different. Finally, each group

was randomly assigned to take different versions of English reading comprehension tests: the ACOM, the ICOM, the ACON and the ICON.

Part II

Descriptive statistics of the reading scores (the dependent variable) obtained from the four versions of English reading comprehension tests (the independent variables) were computed. The mean, median, standard deviation, minimum value, maximum value and range are presented in Table 4.3.

Table 4.3: Descriptive Statistics of the Data

Mediums	Authenticity	Mean	Median	SD	Min	Max	Range
computer	Authentic (ACOM)	17.724	17.33	18.271	11.03	36.23	25.20
	Inauthentic (ICOM)	28.606	23.63	10.425	17.33	55.13	37.80
paper	Authentic (ACON)	23.160	23.00	9.369	3.00	48.00	45.00
	Inauthentic (ICON)	26.387	24.00	9.429	10.00	50.00	40.00

N = 75 for each group

Note: ACOM = Authentic English Reading Comprehension Computer- Adaptive Test
ICOM = Inauthentic English Reading Comprehension Computer- Adaptive Test
ACON = Authentic English Reading Comprehension Conventional Paper-and-Pencil Test
ICON = Inauthentic English Reading Comprehension Conventional Paper-and-Pencil Test

Table 4.3 shows that the highest mean score is the ICOM Group ($\overline{X}$ = 28.606, n= 75), and the lowest mean score is the ACOM group ($\overline{X}$ = 17.724, n = 75).

Checking the assumptions of ANOVA

Before calculating the data in an ANOVA, the underlying assumptions about the nature of the data to be analyzed, and about the robustness of the ANOVA to violations of these assumptions need to be considered. One assumption for the ANOVA is that the populations from which the groups are sampled are normally distributed. A second

assumption is that the variance of the populations from which the groups are drawn are equal. (Bachman, 2004: 245).

Accordingly, the normality of the score distribution and the equality of variance were tested. Table 4.4 presents the results obtained from the normality test.

Table 4.4: Tests of Normality

		Kolmogorov-Smirnov(a)			Shapiro-Wilk		
GROUP		Statistic	df	Sig.	Statistic	df	Sig.
Reading	ACOM	.277	75	.000	.809	75	.000
Scores	ICOM	.217	75	.000	.872	75	.000
	ACON	.168	75	.000	.941	75	.002
	ICON	.129	75	.003	.941	75	.002

$P \leq 0.05$ n = 75 for each group

In order to test the normality of the reading score distribution of each group, Kolmogorov-Smirnov and Shapiro-Wilk were conducted by using SPSS. The Kolmogorov-Smirnov statistics obtained from Group ACOM, ICOM, ACON and ICON were .277, .217, .168 and .129 while Shapiro-Wilk values were .809, .872, .941 and .941 respectively. The significant value from each group was lower than 0.05. It indicated that the distribution of the score from each group was normal (Prap-aripai, 2004: 93).

The homogeneity of variance was tested by Levene Statistic computed by SPSS. The findings are presented in Table 4.5. below.

Table 4.5: Test of Homogeneity of Variance

		Levene Statistic	df1	df2	Sig.
Reading scores	Based on Mean	17.945	3	296	.000
	Based on Median	11.078	3	296	.000
	Based on Median and with adjusted df	11.078	3	238.855	.000
	Based on trimmed mean	16.760	3	296	.000

$P \leq 0.05$

The Based on Mean value calculated by the Levene Statistic was 17.945 and the significant value was .00, which was less than 0.05. This value showed that the variance among the four groups was not equal (Prap-aripai, 2004: 94). However, ANOVA could be conducted in this study since the size of each group was equal (n= 75 for each group). Moreover the ANOVA was reasonably robust to violations of this assumption (Glass & Hopkins, 1996: 405 cited in Bachman 2004: 245).

In conclusion, the data could be analyzed by ANOVA since the score distribution was normal. Although the variance was not equal, the equality of the sizes in each group allowed ANOVA to be robust.

Part III

To investigate the effects of mediums and test authenticity in reading comprehension ability, a 2*2 ANOVA was conducted to find the main effects and interaction effects between two IVs (mediums and authenticity) on DV (reading comprehension score). The data for ANOVA were obtained from the four versions of the English reading comprehension tests: ACOM, ICOM, ACON and ICON. The mean scores from these tests were used in the analyses for main effects and interaction effect. The results were used to test the hypotheses set for this study. The results of the analyses were shown in Table 4.6.

Table 4.6: Tests of Between-Subjects Effects

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	5024.919(a)	3	1674.973	22.064	.000	.183
Intercept	172355.326	1	172355.326	2270.444	.000	.885
Mediums	194.037	1	194.037	2.556	.111	.009
Authenticity	3732.130	1	3732.130	49.164	.000	.142
Mediums * Authenticity	1098.751	1	1098.751	14.474	.000	.047
Error	22470.137	296	75.913			
Total	199850.382	300				
Corrected Total	27495.055	299				

$P \leq 0.05$

The findings were interpreted according to the research questions below:

Research Question 1: Can test authenticity and test delivery mediums have any effects on test takers' English reading proficiency and what are their effect sizes?

Statement of Hypothesis 1: The mean score obtained from the authentic reading comprehension test is significantly different from that obtained from the inauthentic reading comprehension test at the significant level of .05.

Statistical Hypothesis: H1: $\overline{X}$ Authentic $\neq$ $\overline{X}$ Inauthentic

Table 4.6 shows that the significant value of the Authenticity is .00 which is lower than the p value (0.05). According to this, H1 is accepted. This finding indicates that there is significant difference between the mean score obtained from the tests considered authentic (ACOM and ACON) and those obtained from the test considered inauthentic (ICOM and ICON) at the 0.05 level ($p < .05$, $F = 49.164$). This means there is significant difference between the reading scores obtained from the tests with different degrees of authenticity.

The pairwise comparison was conducted by means of Least Significant Difference in order to compare the arithmetic means obtained by the authentic tests (ACOM and ACON) and the inauthentic tests (ICOM and ICON). The findings are presented in Table 4.7.

Table 4.7: Pairwise Comparisons between the Authentic and Inauthentic Tests

					95% Confidence Interval for Difference	
Authenticity	Authenticity	Mean Difference	Std. Error	Sig.(a)	Lower Bound	Upper Bound
authentic	inauthentic	-7.054(*)	1.006	.000	-9.034	-5.074
inauthentic	authentic	7.054(*)	1.006	.000	5.074	9.034

* $P \leq 0.05$

The results obtained from 2*2 ANOVA reveal that there is statistical significance at the level of 0.05. Similarly, the pairwise comparisons computed by Least Significant Difference shows the significant difference between the mean obtained by the authentic

and inauthentic tests (Mean difference is -7.054) at the level of 0.05. This value indicates that the test takers who took the authentic tests had lower reading scores than those who took the inauthentic tests. The mean values for the Authentic Tests (ACOM and ACON) and the Inauthentic Tests (ICOM and ICON) were 20.4420 and 27.4962 respectively.

This effect size was calculated using Partial Eta Squared, which is the typical effect size formula used for univariate design analysis. It is scaled like a percent and interpreted as the higher the number the larger the practical effect. The Partial Eta Squared value can be interpreted according to the following criteria (Cohen,1988 cited in Ary, Jacobs and Razavieh, 2002:360):

>0.2 is large effect size

>0.1 is medium effect size

>0.05 is small effect size

Accordingly, the effect sizes of the two independent variables: test authenticity and test delivery mediums results can be reported by using the finding in Table 4.6.

As illustrated in Table 4.6, the effect size value of authenticity was 142. (See the value in column Partial Eta Squared.) This means that authenticity has medium effect size on test takers' reading ability score (Cohen, 1988 cited in Ary, Jacobs and Razavich, 2002: 360). The medium effect size confirmed the significant difference between the mean obtained by the authentic and inauthentic tests computed by Least Significant Difference.

Hypothesis 2: The mean score obtained from the computer-based reading comprehension test is significantly different from that obtained from the paper-based reading comprehension test at the significant level of .05.

Statistical Hypothesis: H1: $\overline{X}$ CBT $\neq$ $\overline{X}$ PBT

The significant value of the Medium value shown in Table 4.6 is .111, which is higher than p value (0.05). Therefore, the statistical hypothesis is rejected. This finding reveals that there is no significant difference between the mean score obtained from the test delivered by means of computer (ACOM and ICOM) and those obtained from the test delivered by means of paper (ACON and ICON) at the 0.05 level (p> .05, F = .177). This means there is no difference between the reading scores obtained from the tests delivered by different mediums.

Least Significant Difference was computed to conduct the Pairwise comparison between the arithmetic means obtained by the tests delivered by computer (ACOM and ICOM) and paper (ACON and ICON). The findings are shown in Table 4.8.

Table 4.8: Pairwise Comparisons between Tests Delivered by Computer and by Paper

MEDIUMS	MEDIUMS	Mean Difference	Std. Error	Sig.(a)	95% Confidence Interval for Difference Lower Bound	Upper Bound
computer	paper	-1.608	1.006	.111	-3.588	.371
paper	computer	1.608	1.006	.111	-.371	3.588

*P≤ 0.05

The mean difference between computer and paper based tests is -1.608. This value indicates that the test takers who took the paper-based test had higher reading scores than those who took the computer-based test. However, the difference was not statistically significant. The means of the tests delivered by computer (ACOM and ICOM) and the tests delivered by paper (ACON and ICON) were 23.1649 and 24.7733 respectively.

The effect size value illustrated in Table 4.6 of test delivery mediums is .009. (See the values in column Partial Eta Squared.) This means that the effect size of the mediums on test takers' reading ability scores was small (Cohen, 1988 cited in Ary, Jacobs and Razavieh, 2002: 360).

Hypothesis 3: There are significant interaction effects between test authenticity and test delivery mediums on students' reading proficiency at the significant level of .05.

Statistical Hypotheses:

H 3.1: $\overline{X}$ Authentic CBT ≠ $\overline{X}$ Inauthentic CBT

H 3.2: $\overline{X}$ Authentic PBT ≠ $\overline{X}$ Inauthentic PBT

H 3.3: $\overline{X}$ Authentic CBT ≠ $\overline{X}$ Authentic PBT

H 3.4: $\overline{X}$ Inauthentic CBT ≠ $\overline{X}$ Inauthentic PBT

H 3.5: $\overline{X}$ Authentic CBT ≠ $\overline{X}$ Inauthentic PBT

H 3.6: $\overline{X}$ Inauthentic CBT ≠ $\overline{X}$ Authentic PBT

According to the results presented in Table 4.6, the interaction effect between Mediums and Authenticity was found at the 0.05 significant level ($p<.05$, $F = 14.474$). The plotted graph of the interaction effect between test delivery mediums and test authenticity can be seen in Figure 4.1.

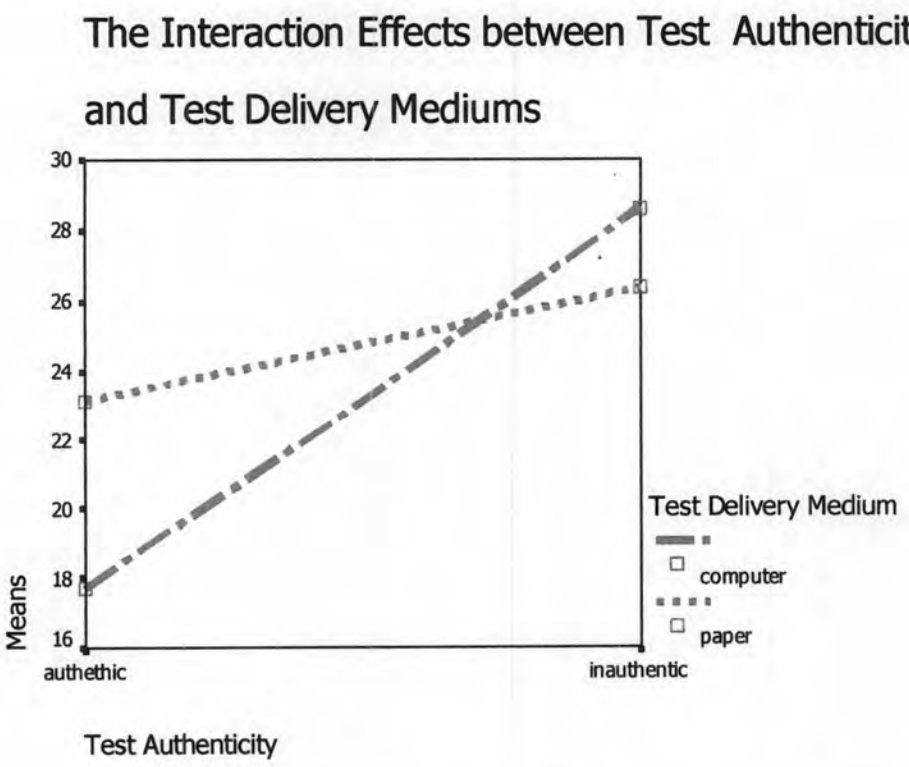

Figure 4.1: Plot of the interaction effect between mediums and test authenticity

Figure 4.1 shows the plotted graph of the mean scores from the tests possessing different delivery mediums and degree of test authenticity. The disordinal lines illustrate there is a significant interaction effect between mediums and authenticity.

It can be interpreted that the high degree of test authenticity causes lower scores in the tests delivered by the computer. Conversely, the low degree of test authenticity or the tests considered inauthentic in this study allows higher scores in the test delivered by the computer.

In terms of the tests delivered by paper, the test takers have the tendency to gain higher scores when they take the tests with a low degree of test authenticity (the inauthentic tests). Similarly, the test takers who take the tests delivered by computer tend to have higher scores when the tests are less authentic.

Part IV

Research Question 2: What are the test takers' attitudes towards test authenticity and test delivery mediums?

The interview was conducted with five students randomly selected from each group (total 20 students from 4 groups) in their first language (Thai). Then the data were recorded on video. Once gathered, the video recordings were transcribed.

An encoding scheme (see Appendix 19) was used for the transcribed script to facilitate analysis. The test takers' responses were encoded as positive, negative and neutral attitudes.

In order to ensure the reliability of the encoding, inter-coder reliability was computed to assess the extent to which the two coders agreed on the codes assigned to each segment. A high level of agreement (80 per cent and above) is usually sought between coders (Green, 2004).

In order to prepare the data for computing the inter-coder reliability, three scales were assigned to each encoded attitude as follows:

Positive attitude = +1

Neutral attitude = 0

Negative attitude = -1

The coders' judgments on the test takers' responses in each main segment were summed up (see Appendix 20). In total, the number of all responses from twenty test takers was 80. Then the summed up scores were computed by means of Pearson Correlation to investigate the inter-coder reliability. Correlations between the two coders are presented in Table 4.9.

Table 4.9: Correlations between two coders

Coders		**Coder 1**	**Coder 2**
Coder 1	Pearson Correlation	1	.901(**)
	Sig. (2-tailed)	.	.000
	From N	80	80
Coder 2	Pearson Correlation	.901(**)	1
	Sig. (2-tailed)	.000	.
	From N	80	80

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

From Table 4.9, it can be seen that in encoding the 80 qualitative responses from 20 test takers (5 from each group), the correlation between Coder 1 and Coder 2 was .901. The correlation is significant at the 0.01 level (2-tailed).

In summary, in terms of inter-coder reliability, high correlation (> 80%) between two coders was found. This indicates the degree to which the encoding of one coder can be predicted from the encodings of the other coders (Hatch and Farhady, 1982: 203).

In order to reveal test takers' attitudes, a content analysis technique was employed. The approach to content analysis used in this study was by frequency counts. In this approach, the units for coding identified and coding categories (positive, neutral and negative attitudes) defined were tabulated and then carefully counted.

Moreover, the chi square (χ^2) test was applied to find out whether there was any significant difference in the proportions of samples' attitudes towards test authenticity and test delivery among the four groups. However, Isaac and Michael (1983: 177) mentioned one restriction on the use of chi square that no theoretical frequency should be smaller than 5. Because the number of test takers randomly selected from each group was 5, there was a high possibility that the number was likely to be less than 5. The Exact Tests in SPSS can be used when the number is smaller than 5 (Silpjaru, 2005: 374). This study, therefore, conducted the chi square by means of the Exact Tests to compare the frequencies of the four responding groups.

Accordingly, the findings were reported according to the topics and issues to be covered in the semi-structured interview as follows:

1.General questions asking about the quality of the test to assess reading ability

1.1 *The opportunity to demonstrate strengths and weaknesses in reading ability*

Question: Do you think that you have a sufficient opportunity in the test to demonstrate your strengths and weaknesses in reading ability?

The frequencies of all test takers' responses were tested by chi square. The results were presented in Table 4.10.

Table 4.10: Chi square Tests of the Test Takers' Opinion on the Opportunity to Demonstrate Strengths and Weaknesses in Reading Ability

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	3.158(a)	3	.368	1.000		
Likelihood Ratio	2.937	3	.402	1.000		
Fisher's Exact Test	2.958			1.000		
Linear-by-Linear Association	1.800(b)	1	.180	.500	.250	.250
N of Valid Cases	20					

$P \leq 0.05$

The significant value of 0.368 indicated that the frequencies of the four groups were not significantly different at the level of 0.05.

Table 4.11 illustrates the frequency and percentage of the test takers' opinion on the opportunity to demonstrate strengths and weaknesses in reading ability from the four groups.

Table 4.11: Frequency and Percentage of the Test Takers' Opinions on the Opportunity to Demonstrate Strengths and Weaknesses in Reading Ability

Investigated topics	Frequency and Percentage	Groups			
		ACOM	ICOM	ACON	ICON
Strength and weakness					
(positive)	Frequency	5	5	5	4
	Expected frequency	4.8	4.8	4.8	4.8
	Percentage	100	100	100	80
(neutral)	Frequency	0	0	0	0
	Expected frequency	0.3	0.3	0.3	0.3
	Percentage	0	0	0	20
(negative)	Frequency	0	0	0	1
	Expected frequency	0.3	0.3	0.3	0.3
	Percentage	0	0	0	20

N=5 for each group

Table 4.11 shows that 100 per cent of Group ACOM, ICOM, ACON and 80 per cent of Group ICON positively agreed that the tests provided them the opportunity to demonstrate strengths and weaknesses in reading ability.

Qualitative responses: the test takers in all four groups selected positive statements on the opportunity to demonstrate strengths and weaknesses in reading ability as follows:

After taking the test, I realize what level of my reading ability is.

One big problem of my reading skill is to find out the main idea, especially in the long paragraph. Anyway, I feel that I am quite good at reading the paragraph relevant to the job classified.

I know that there are lots of vocabulary I don't know. Because of this, I can't answer many questions in the test.

The test lets me know my weaknesses so that I realize what I should improve.

1.2 *The perception of test difficulty*

Question: How difficult is the test?

Table 4.12: Chi-Square Tests of the Test Takers' Opinion on the Perception of Test Difficulty

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	3.222(a)	6	.780	.883		
Likelihood Ratio	4.132	6	.659	.883		
Fisher's Exact Test	3.809			.883		
Linear-by-Linear Association	1.356(b)	1	.244	.296	.148	.044
N of Valid Cases	20					

$P \leq 0.05$

Table 4.12 shows that the significant value was 0.780 which indicates that the frequencies of the four groups were not significantly different at the level of 0.05. This can be interpreted as undifferentiated test takers' perceptions of test difficulty. Table 4.13 illustrates the frequency and percentage of the test takers' perception of test difficulty from the four groups.

Table 4.13: Frequency and Percentage of the Test Takers' Perception of Test Difficulty

Investigated topics	Frequency and Percentage	Groups			
		ACOM	ICOM	ACON	ICON
Test Difficulty					
(positive)	Frequency	0	0	1	1
	Expected frequency	0.5	0.5	0.5	0.5
	Percentage	0	0	20	20
(neutral)	Frequency	3	3	3	3
	Expected frequency	2.7	2.7	2.7	2.7
	Percentage	60	60	60	60
(negative)	Frequency	2	2	1	1
	Expected frequency	1	1.5	1.5	1.5
	Percentage	40	40	20	20

N=5 for each group

Table 4.13 shows that all four groups have the same frequencies on the neutral perception of the test difficulty. This can be interpreted as test difficulty being perceived as neutral.

Qualitative responses: The neutral opinions on the test difficulty from all four groups are summarized as follows:

I perceive that the test difficulty is just medium; it is not too hard and not too easy.

It is not too hard. The questions may not be difficult, but without any choices some items are.

If we pay attention to the paragraphs, we can do the test. It is not too hard.

1.3 *The perception of test fairness*

Question: How do you feel about the questions in the test? Do they affect your score? How do you feel about the situations in the test? Do they affect your score?

The test takers' responses can be sub-divided into the perception of test fairness on:

1.3.1 the appropriateness of the questions,

1.3.2 the effect of examinee response methods and

1.3.3 the situations in the tests.

1.3.1 The appropriateness of the questions

Table 4.14 shows the results calculated by chi square on test fairness focusing on opinions on the appropriateness of the questions.

Table 4.14: Chi-Square Tests of the Test Takers' Perceptions of Test Fairness: the Appropriateness of the Questions

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	2.286(a)	6	.892	1.000		
Likelihood Ratio	3.739	6	.712	1.000		
Fisher's Exact Test	3.243			1.000		
Linear-by-Linear Association	.017(b)	1	.895	1.000	.500	.100
N of Valid Cases	20					

$P \leq 0.05$

The significant value presented in Table 4.14 is 0.892, which indicates that the frequencies of the four groups were not significantly different at the level of 0.05. The interpretation is that the test takers' perceptions of the appropriateness of the questions are not different. Table 4.15 shows the frequency and percentage of the test takers' perceptions of test fairness: the appropriateness of the questions obtained from the four groups.

Table 4.15: Frequency and Percentage of the Test Takers' Perceptions of Test Fairness: the Appropriateness of the Questions

Investigated topics	Frequency and Percentage	Groups			
		ACOM	ICOM	ACON	ICON
Test Fairness: questions					
(positive)	Frequency	4	3	4	3
	Expected frequency	3.5	3.5	3.5	3.5
	Percentage	80	60	80	60
(neutral)	Frequency	0	1	1	1
	Expected frequency	0.8	0.8	0.8	0.8
	Percentage	0	20	20	20
(negative)	Frequency	1	1	0	1
	Expected frequency	0.8	0.8	0.8	0.8
	Percentage	20	20	0	20

N=5 for each group

Table 4.15 shows that the majority of Group ACOM, ICOM, ACON and ICON positively agreed that the questions in the test were appropriate.

Qualitative responses: the test takers from the four groups reported the positive statements to the appropriateness of the questions as follows:

They are easy to understand, not complicated.

If I have done the item wrong, it is because of my own mistake. It is not the result of the complexity of the questions.

1.3.2 The effect of the ways to answer the tests

Table 4.16 illustrates Chi-Square tests of the test takers' perceptions of test fairness: the effect of the ways responding to the tests.

Table 4.16: Chi-Square Tests of the Test Takers' Perceptions of Test Fairness: the Effect of Examinee Response Methods

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	3.222(a)	6	.780	.883		
Likelihood Ratio	4.132	6	.659	.883		
Fisher's Exact Test	3.809			.883		
Linear-by-Linear Association	1.356(b)	1	.244	.296	.148	.044
N of Valid Cases	20					

$P \leq 0.05$

Table 4.16 shows that the significant value is 0.780, which indicates that the frequencies of the four groups are not significantly different at the level of 0.05. It can be interpreted that the test takers' perceptions of the ways to answer the tests are not statistically different.

The following table demonstrates the frequency and percentage of the test takers' perceptions of test fairness: the effect of the ways to respond in the tests.

Table 4.17: Frequency and Percentage of the Test Takers' Perceptions of Test Fairness: the Effect of the Ways to Answer the Tests

Investigated topics	**Frequency and Percentage**	**Groups**			
		ACOM	**ICOM**	**ACON**	**ICON**
Test Fairness: responses					
(positive)	Frequency	1	2	3	3
	Expected frequency	2.3	2.3	2.3	2.3
	Percentage	20	40	60	60
(neutral)	Frequency	1	1	1	0
	Expected frequency	0.8	0.8	0.8	0.8
	Percentage	20	20	20	0
(negative)	Frequency	3	2	1	2
	Expected frequency	2	2	2	2
	Percentage	60	40	20	40

N=5 for each group

Table 14.7 shows that Group ACOM, ICOM, ACON and ICON have similar frequencies in the negative opinions on the effect of the ways responding to the tests. However, Group ACOM has the highest frequencies (f=3), while ICOM and ICON have the same frequencies (f=2) and ACON has only one negative frequency.

Qualitative responses: the negative statements to the effects of the ways to answer the test are summarized as follows:

After typing, if I did not recheck, I may lose the score. (ACOM)

The choices in each item may fool me when I do not understand the text. (ICOM and ICON).

I am worried about my grammar when I write the answers. (ACON)

1.3.3 The situations in the tests.

Following are the statistical results of Chi-Square tests of the test takers' perceptions of test fairness: the situations in the tests.

Table 4.18: Chi-Square Tests of the Test Takers' Perceptions of Test Fairness: the Situations in the Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	13.429(a)	6	.037	.027		
Likelihood Ratio	16.211	6	.013	.035		
Fisher's Exact Test	11.535			.028		
Linear-by-Linear Association	6.927(b)	1	.008	.009	.005	.002
N of Valid Cases	20					

$P \leq 0.05$

Table 4.18 shows that the significant value is 0.037 which indicates that the frequencies of the four groups were significantly different at the level of .05. It can be interpreted that the test takers' perceptions of the situations in the tests affecting their scores were significantly different at .05 level. Table 4.19 shows the frequency and percentage of the test takers' perceptions of test fairness: the situations in the tests.

Table 4.19: Frequency and Percentage of the Test Takers' Perceptions of Test Fairness: the Situations in the Tests

Investigated topics	Frequency and Percentage	Groups			
		ACOM	ICOM	ACON	ICON
Test Fairness: situations					
(positive)	Frequency	3	3	3	0
	Expected frequency	2.3	2.3	2.3	2.3
	Percentage	60	60	60	0
(neutral)	Frequency	2	1	1	0
	Expected frequency	1	1	1	0
	Percentage	40	20	20	0
(negative)	Frequency	0	1	1	5
	Expected frequency	1.8	1.8	1.8	1.8
	Percentage	0	20	20	100

N=5 for each group

Table 4.19 shows that the three groups: Group ACOM, ICOM and ACON positively agreed that the test situations affected their scores. Conversely, 100 percent of Group ICON selected a negative opinion for the effect of the situations in the tests.

Due to the significant difference, the bar graph below illustrates the differences among the four groups.

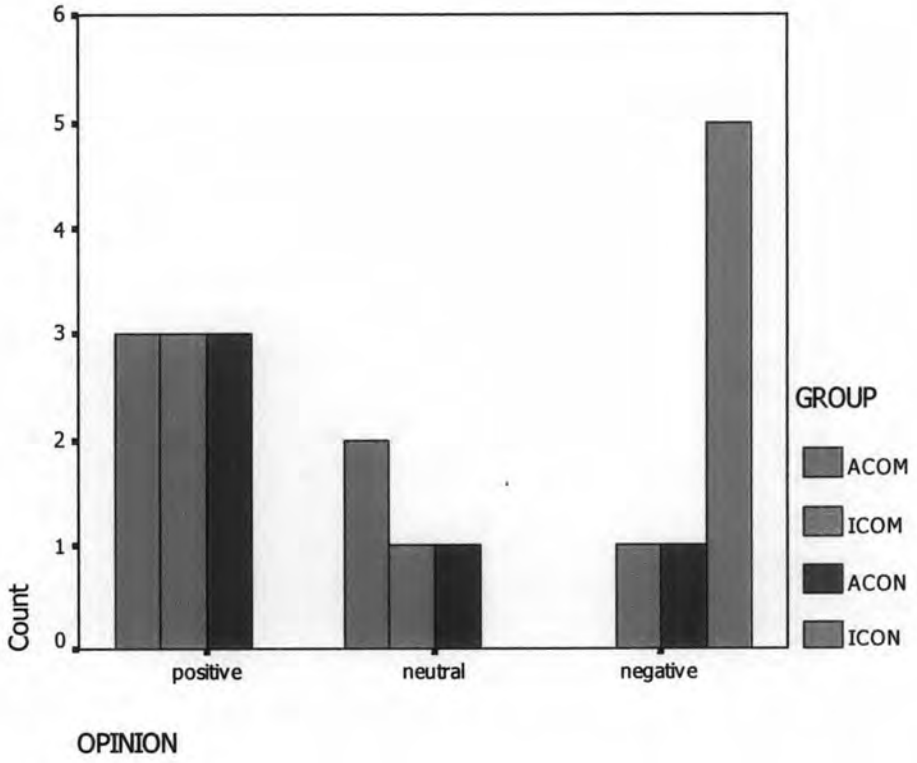

Figure 4.2: Difference frequencies among the four groups in terms of the test takers' perceptions of test fairness: the situations in the tests

Qualitative responses: Group ACOM, ICOM and ACON selected positive statements to describe the effects of the situations in the tests as follows:

The situations in the test are appropriate. I can imagine that I were that character and tried to answer the questions as I were in that situations.

The situations allow me to understand the text more.

Qualitative responses: However, 100 per cent of the ICON group selected negative opinions as follows:

The multiple choice items are unusual tasks that occur in real life. Anyway, I don't think there is such situation in the test.

I am not sure about the situations in the test. If there are, they might not help increase the score.

1.4 *The perception of nervousness while taking the test*

Question: How do you feel while taking the test? Are you in control of the test situation or do you feel nervous?

Table 4.20 illustrates Chi-Square tests of the test takers' perception of nervousness while taking the tests.

Table 4.20: Chi-Square Tests of the Test Takers' Perception of Nervousness While Taking the Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	7.727(a)	6	.259	.286		
Likelihood Ratio	7.564	6	.272	.365		
Fisher's Exact Test	7.041			.365		
Linear-by-Linear Association	1.239(b)	1	.266	.319	.159	.045
N of Valid Cases	20					

$P \leq 0.05$

Table 4.20 shows that the significant value is 0.259 which indicates that the frequencies of the four groups are not significantly different at the level of 0.05. It can be interpreted that the test takers' perceptions of nervousness in the four groups while taking the test are not statistically different.

Table 4.21 points out the frequency and percentage of the test takers' perception of nervousness while taking the test.

Table 4.21: Frequency and Percentage of the Test Takers' Perception of Nervousness While Taking the Test

Investigated topics	Frequency and Percentage	Groups			
		ACOM	ICOM	ACON	ICON
Nervousness					
(positive)	Frequency	3	1	3	4
	Expected frequency	2.8	2.8	2.8	2.8
	Percentage	60	20	60	80
(neutral)	Frequency	0	0	1	0
	Expected frequency	0.3	0.3	0.3	0.3
	Percentage	0	0	20	0
(negative)	Frequency	2	4	1	1
	Expected frequency	2	2	2	2
	Percentage	40	80	20	20

N=5 for each group

Table 4.21 shows that the majority of Group ACOM (f=3), ACON (f=3) and ICON (f=4) positively agree on the perception of nervousness while taking the test. However, only the majority of Group ICOM (f=4) gave a negative opinion on the perception of nervousness.

Qualitative responses: The positive statements made by the test takers from the four groups are summarized as follows:

I felt as if I did ordinary exercises.

I was not excited at all, and felt fun when the text topic was my interest.

1.5 *The perception of the clarity of the test directions*

Question: How clear are the test directions? Do you know what to do in the test?

All the responses from the four groups were constant. All provided the same opinions, so the Chi Square was not computed. Table 4.22 shows the frequency and percentage of the test takers' perception of the clarity of test directions.

Table 4.22: Frequency and Percentage of the Test Takers' Perception of the Clarity of the Test Directions

Investigated topics	Frequency and Percentage	Groups			
		ACOM	ICOM	ACON	ICON
Test direction					
(positive)	Frequency	5	5	5	5
	Expected frequency	5	5	5	5
	Percentage	100	100	100	100
(neutral)	Frequency	0	0	0	0
	Expected frequency	0.3	0.3	0.3	0.3
	Percentage	0	0	0	0
(negative)	Frequency	0	0	0	0
	Expected frequency	0.3	0.3	0.3	0.3
	Percentage	0	0	0	0

N=5 for each group

Table 4.22 shows that 100 per cent of Group ACOM, ICOM, ACON and ICON positively agreed that the test directions were clear.

Qualitative responses: The positive statements which indicate that they all agreed on the clarity of the test directions are summarized as follows:

The instructions are clear. I know what I will do in the test.

An example of the test is provided in the program, so I know what the test is like before taking it.

1.6 *The accuracy of the test to elicit their true ability in reading*

Question: How accurately does the test elicit your true reading ability? Does your true reading ability show in your real life reading activities?

The responses can be categorized into the following items:

1.6.1 accuracy in eliciting true ability in reading, and

1.6.2 the relevance of the test tasks to real life reading activities

1.6.1 accuracy in eliciting true ability in reading

Table 4.23 illustrates Chi-Square tests of the test takers' opinions on the accuracy of the test to elicit their true ability in reading.

Table 4.23: Chi-Square Tests of the Test Takers' Opinions on the Accuracy of the Test to Elicit Their True Ability in Reading

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	5.176(a)	6	.521	1.000		
Likelihood Ratio	5.715	6	.456	1.000		
Fisher's Exact Test	5.089			1.000		
Linear-by-Linear Association	.146(b)	1	.702	.883	.442	.139
N of Valid Cases	20					

$P \leq 0.05$

Table 4.23 shows that the significant value is 0.521 which indicates that the frequencies of the four groups were not significantly different at the level of 0.05. It can be interpreted that the test takers' opinions on the accuracy of the test to elicit their true ability in reading are not statistically different. Table 4.24 illustrates the frequency and

percentage of the test takers' opinions on the accuracy of the test to elicit their true ability in reading: the accuracy to elicit their true ability in reading.

Table 4.24: Frequency and Percentage of the Test Takers' Opinions on the Accuracy of the Test to Elicit their True Ability in Reading: the Accuracy to Elicit Their True Ability in Reading

Investigated topics	Frequency and Percentage	Groups			
		ACOM	ICOM	ACON	ICON
Accuracy					
(positive)	Frequency	4	4	5	4
	Expected frequency	4.3	4.3	4.3	4.3
	Percentage	80	80	100	80
(neutral)	Frequency	1	1	0	0
	Expected frequency	0.5	0.5	0.5	0.5
	Percentage	20	20	0	0
(negative)	Frequency	0	0	0	1
	Expected frequency	0.3	0.3	0.3	0.3
	Percentage	0	0	0	20

N=5 for each group

Table 4.24 shows that the majority of Group ACOM (f=4), ICOM (f=4), ACON (f=5) and ICON (f=4) positively agreed that the tests elicited their true ability in reading accurately.

Qualitative responses: The positive statements are summarized as follows:

Absolutely accurate, the ones I have wrong come from the lack of my knowledge.

1.6.2 There is relevance of the test tasks to real life reading activities.

Table 4.25 Chi-Square tests of the test takers' opinions on the accuracy of the test to elicit their true ability in reading: the relevance of the test tasks to their real life reading activities.

Table 4.25: Chi-Square Tests of the Test Takers' Opinions on the Accuracy of the Test to Elicit Their True Ability in Reading: the Relevance of the Test Tasks to Their Real Life Reading Activities

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	7.500(a)	3	.058	.128		
Likelihood Ratio	8.282	3	.041	.128		
Fisher's Exact Test	5.689			.128		
Linear-by-Linear Association	3.800(b)	1	.051	.085	.042	.031
N of Valid Cases	20					

$P \leq 0.05$

Table 4.25 shows that the significant value is 0.058 which indicates that the frequencies of the four groups are not significantly different at the level of 0.05. It can be interpreted that the test takers' opinions on the relevance of the test tasks to their real life reading activities are not significantly different.

Table 4.26 shows the frequency and percentage of the test takers' opinions on the accuracy of the test to elicit their true ability in reading: the relevance of the test tasks to their real life reading activities.

Table 4.26: Frequency and Percentage of the Test Takers' Opinions on the Accuracy of the Test to Elicit Their True Ability in Reading: the Relevance of the Test Tasks to Their Real Life Reading Activities

Investigated topics	Frequency and Percentage	Groups			
		ACOM	ICOM	ACON	ICON
Real life activities					
(positive)	Frequency	5	4	5	2
	Expected frequency	4	4	4	4
	Percentage	100	80	100	40
(neutral)	Frequency	0	1	0	3
	Expected frequency	1	1	1	1
	Percentage	0	20	0	60
(negative)	Frequency	0	0	0	0
	Expected frequency	0.3	0.3	0.3	0.3
	Percentage	0	0	0	0

N=5 for each group

Table 4.26 shows that the majority of Group ACOM (f=5), ICOM (f=4) , ACON (f=5) positively agree that the tests can elicit their true ability because the test tasks are relevant to their real life reading activities. However, only 2 respondents from ICON reported positive opinion.

Qualitative responses: The positive statements which indicate that the majority agree that the tests can elicit their true ability because the test tasks are relevant to their real life reading activities are summarized as follows:

They are relevant to my real life reading activities because I normally read job classified and English news.

When I talk to foreigners I have to organize sentences or phrases similarly to what I did in the test.

Responding to the short answer items is similar to discussing on the web board.

2 The test administration:

2.1 *Test administration*:

Question: How do you like the test administration? Do you think it is more convenient than the traditional test?

The responses are divided into two main parts as follows:

2.1 test administration: the perception on the test administration they took, and

2.2 test administration compared to other versions of the tests.

2.1 Test administration

Table 4.27 illustrates Chi-Square tests of the test takers' opinions on the test administration.

Table 4.27: Chi-Square Tests of the Test Takers' Opinions on the Test Administration: The Perception on the Test Administration They Took

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	5.524(a)	6	.479	.649		
Likelihood Ratio	7.191	6	.304	.649		
Fisher's Exact Test	5.204			.649		
Linear-by-Linear Association	.017(b)	1	.895	1.000	.500	.100
N of Valid Cases	20					

$P \leq 0.05$

Table 4.27 shows that the significant value is 0.479 which indicates that the frequencies of the four groups are not significantly different at the level of 0.05. It can be interpreted that the test takers' opinions on the test administration were not significantly different.

Table 4.28 shows the frequency and percentage of the test takers' opinions on the test administration.

Table 4.28: Frequency and Percentage of the Test Takers' Opinions on the Test Administration: The Perception on the Test Administration They Took

Investigated topics	Frequency and Percentage	Groups			
		ACOM	ICOM	ACON	ICON
Test administration					
(positive)	Frequency	4	4	2	4
	Expected frequency	3.5	3.5	3.5	3.5
	Percentage	80	80	40	80
(neutral)	Frequency	0	0	2	1
	Expected frequency	0.8	0.8	0.8	0.8
	Percentage	0	0	40	20
(negative)	Frequency	1	1	1	0
	Expected frequency	0.8	0.8	0.8	0.8
	Percentage	20	20	20	0

N=5 for each group

Table 4.28 shows the majority of Group ACOM (f=4), ICOM (f=4), and ICON (f=4) positively agrees on the test administration. However, there were two respondents from ACON who had positive opinions.

Qualitative responses: The positive statements selected by the majority agreeing with test administration are as follows:

I like this kind of administration because it is convenient and not complicated.

2.2 Test administration compared to other versions of the tests

The following table presents Chi-Square tests of the test takers' opinions on test administration: comparison with other versions of the tests.

Table 4.29: Chi-Square Tests of the Test Takers' Opinions on the Test Administration: Compared to Other Versions of the Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	5.429(a)	6	.490	.530		
Likelihood Ratio	7.017	6	.319	.505		
Fisher's Exact Test	5.428			.505		
Linear-by-Linear Association	.935(b)	1	.333	.413	.207	.065
N of Valid Cases	20					

$P \leq 0.05$

Table 4.29 shows that the significant value is 0.490 which indicates that the frequencies of the four groups are not significantly different at the level of 0.05. It can be interpreted that the test takers' opinions on test administration when compared to other versions of the tests are not significantly different.

Table 4.30 illustrates the frequency and percentage of the test takers' opinions on test administration: test administration compared to other versions of the tests.

Table 4.30: Frequency and Percentage of the Test Takers' Opinions on the Test Administration: Test Administration Compared to Other Versions of the Tests

Investigated topics	Frequency and Percentage	Groups			
		ACOM	ICOM	ACON	ICON
Compared between paper and computer versions					
(positive)	Frequency	4	2	3	5
	Expected frequency	3.5	3.5	3.5	3.5
	Percentage	80	40	60	100
(neutral)	Frequency	0	1	1	0
	Expected frequency	0.5	0.5	0.5	0.5
	Percentage	0	20	20	0
(negative)	Frequency	1	2	1	0
	Expected frequency	1	1	1	1
	Percentage	20	40	20	20

N=5 for each group

Table 4.30 shows that the majority of Group ACOM (f=4), ACON (f=3) and ICON (f=5) positively agrees on the test administration when compared to other versions of the tests. The results are not statistically different. However, two respondents from Group ICOM have the following positive opinions.

Qualitative responses: The positive statements which indicated that the majority preferred the test administration they took rather than other versions are summarized as follows:

Students who took the paper-based tests compared computer-based tests (ACON and ICON) with the following statements: This paper-based test is more appropriate. I think it is more convenient because we can take it any time and any place. We can go back to the previous items any time.

Students who took the computer-based tests compared them with paper-based tests (ACOM and ICOM) with the following statements: The computer test is more convenient. Paper-based tests may use a lot of paper, but this test administration can reduce using paper. We immediately know the result when we finish the test.

2.2 *Time length*

Question: How sufficient is the time provided for taking the test? Are you satisfied with the length of the time used in the test?

The following table presents Chi-Square tests of the test takers' opinions on the time length.

Table 4.31: Chi-Square Tests of the Test Takers' Opinions on the Time Length

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	15.000(a)	3	.002	.004		
Likelihood Ratio	15.012	3	.002	.004		
Fisher's Exact Test	10.295			.004		
Linear-by-Linear Association	.950(b)	1	.330	.479	.239	.125
N of Valid Cases	20					

$P \leq 0.05$

Table 4.31 shows that the significant value is 0.002 which indicates that the frequencies of the four groups are significantly different at the level of 0.05. It can be interpreted that the test takers' opinions on the time length are significantly different.

Table 4.32 shows the frequency and percentage of the test takers' opinions on the time length.

Table 4.32: Frequency and Percentage of the Test Takers' Opinions on the Time Length

Investigated topics	Frequency and Percentage	Groups			
		ACOM	ICOM	ACON	ICON
Time length					
(positive)	Frequency	5	5	1	5
	Expected frequency	4	4	4	4
	Percentage	100	100	20	100
(neutral)	Frequency	0	0	0	0
	Expected frequency	0.3	0.3	0.3	0.3
	Percentage	0	0	0	0
(negative)	Frequency	0	0	4	0
	Expected frequency	1	1	1	1
	Percentage	0	0	80	0

N=5 for each group

Table 4.32 shows that 100 per cent of Group ACOM (f=5), ICOM (f=5) and ICON (f=5) positively agree on time length. Conversely, the majority of Group ACON has negative opinions which are described below.

The bar graph below presents the differences compared among the four groups.

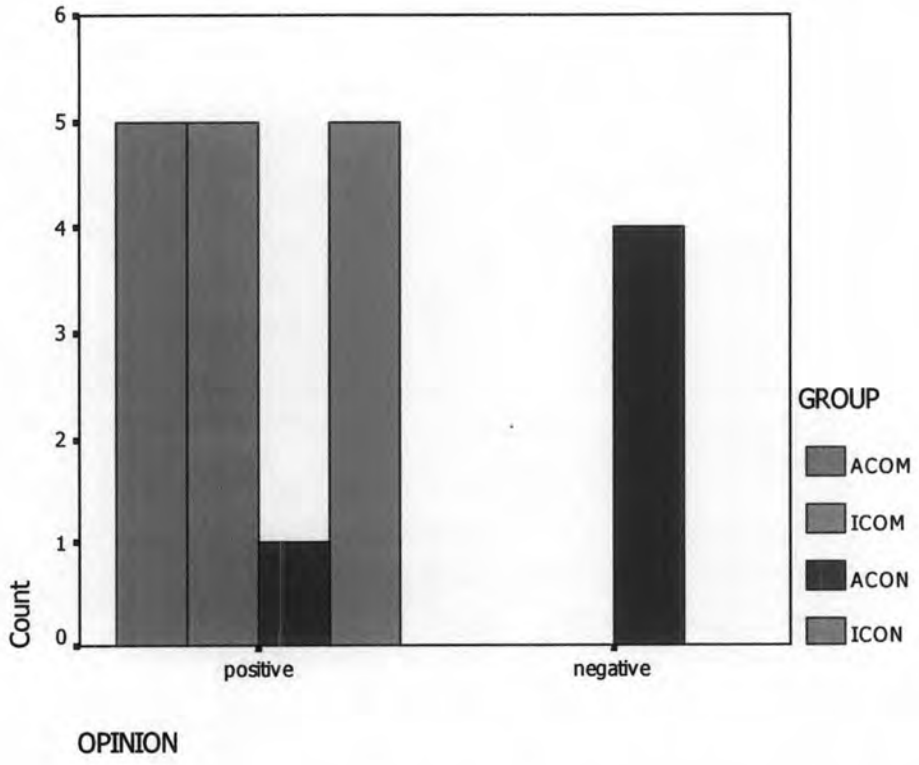

Figure 4.3: Difference frequencies among the four groups in terms of the test takers' opinions on the time length

Qualitative responses: Group ACOM, ICOM and ICON reported the positive statements to the effects of the time length as follows:

(ICON): I have plenty of time to finish the test. Two hours are sufficient.

(ACOM and ICOM): I was not in a rush to finish the test. I spent only 45 to one hour to finish the test. I am really satisfied with the time length spent.

Qualitative responses: However, the majority of ACON reported the negative opinion. Their opinions can be summarized as follows:

Only two hours are not sufficient. I could not finish the test in time. It would be better if we were allowed to have one more hour.

2.3 *Facilities*

When asked about facilities, the results were as follows:

Question: How effective are the facilities used in each test administration (PBT and CBT)? In terms of CBT, how do you like the software functions?

Table 4.33 illustrates Chi-Square tests of the test takers' opinions on the facilities.

Table 4.33: Chi-Square Tests of the Test Takers' Opinions on the Facilities

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	5.647(a)	6	.464	.561		
Likelihood Ratio	6.221	6	.399	.561		
Fisher's Exact Test	5.536			.561		
Linear-by-Linear Association	.221(b)	1	.639	.766	.383	.110
N of Valid Cases	20					

$P \leq 0.05$

Table 4.33 shows that the significant value is 0.464 which indicates that the frequencies of the four groups are not significantly different at the level of 0.05. It can be interpreted that the test takers' opinions on the facilities are not significantly different.

Table 4.34 shows the frequency and percentage of the test takers' opinions on the facilities.

Table 4.34: Frequency and Percentage of Test Takers' Opinions on the Facilities

Investigated topics	Frequency and Percentage	Groups			
		ACOM	ICOM	ACON	ICON
Facilities					
(positive)	Frequency	4	5	3	5
	Expected frequency	4.3	4.3	4.3	4.3
	Percentage	80	100	60	100
(neutral)	Frequency	0	0	1	0
	Expected frequency	0.3	0.3	0.3	0.3
	Percentage	0	0	20	0
(negative)	Frequency	1	0	1	0
	Expected frequency	0.5	0.5	0.5	0.5
	Percentage	20	0	20	0

N=5 for each group

Table 4.34 shows that the majority of Group ACOM, ICOM, ACON and ICON positively agree on the facilities used in the tests. However, the ACON group has different opinions. Following are their responses.

Qualitative responses: The positive statements which reveal their agreement on the facilities used in the tests are summarized as follows:

(ACON and ICON): All the facilities in the exam room is OK. They allow us to take the test without any problems occurring during the test.

(ACOM and ICOM): It a good software. The system is standard so that the results obtained are accurate. One important advantage is that it reports the weaknesses. We know which items we have right or wrong after taking the test.

3 The test characteristics

This part covers the students' opinions on test contents, scoring, interactiveness, authenticity and familiarity.

3.1. *Test contents*

Question: How interesting are the test contents?

Table 4.35 shows Chi-Square tests of the test takers' opinion on the test contents.

Table 4.35: Chi-Square Tests of the Test Takers' Opinion on the Test Contents

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	10.588(a)	3	.014	.035		
Likelihood Ratio	10.178	3	.017	.035		
Fisher's Exact Test	6.696			.035		
Linear-by-Linear Association	6.035(b)	1	.014	.018	.009	.009
N of Valid Cases	20					

$P \leq 0.05$

Table 4.35 shows that the significant value is 0.014 which indicates that the frequencies of the four groups are significantly different at the level of 0.05. It can be interpreted that the test takers' opinion on the test contents were significantly different.

Table 4.36 presents the frequency and percentage of the test takers' opinion on the test contents.

Table 4.36: Frequency and Percentage of the Test Takers' Opinion on the Test Contents

Investigated topics	Frequency and Percentage	Groups			
		ACOM	ICOM	ACON	ICON
Test content					
(positive)	Frequency	5	5	5	2
	Expected frequency	4.3	4.3	4.3	4.3
	Percentage	100	100	100	40
(neutral)	Frequency	0	0	0	3
	Expected frequency	0.3	0.3	0.3	0.3
	Percentage	0	0	0	60
(negative)	Frequency	0	0	0	0
	Expected frequency	0.3	0.3	0.3	0.3
	Percentage	0	0	0	0

N=5 for each group

Table 4.36 shows that the majority of Group ACOM (f=5), ICOM (f=5), ACON (f=5) positively agreed on the interest of the test content while the majority of Group ICON (f=3) neutrally agreed on the interest of the test content.

The following bar chart illustrates the differences among the four groups.

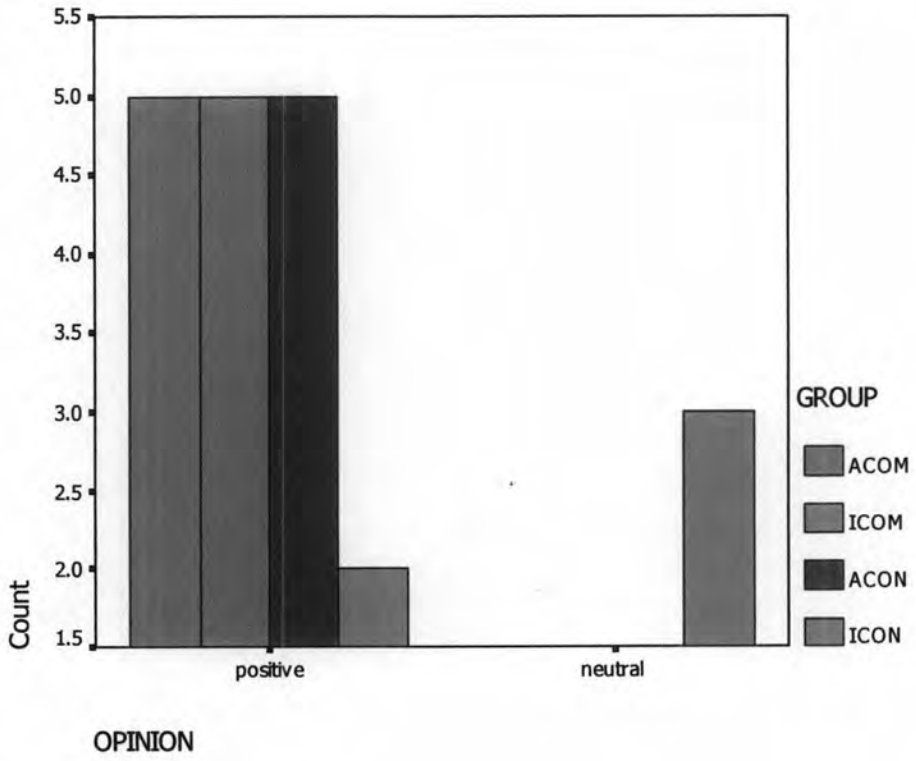

Figure 4.4: Difference frequencies among the four groups in terms of the test takers' opinion on the test contents

Qualitative responses: Group ACOM, ICOM and ACON selected the positive statements to the effects of the situations in the tests as follows:

The reading texts are very interesting, especially the last one which is about GPS. It is relevant to what I am studying.

All the texts are relevant to our everyday life. Normally we like to read entertainment news.

The text contents go well with the situation content in the test. For example, after reading the text we use what we read to act as that character who tries to communicate with another character in the test items.

However, the majority of Group ICON selected the neutral opinion. They felt that the test was similar to ordinary tests.

3.2 *Scoring*

Question: How effective is the scoring method?

Table 4.37 illustrates Chi-Square tests of the test takers' opinions on the scoring.

Table 4.37: Chi-Square Tests of the Test Takers' Opinions on the Scoring

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	11.667(a)	3	.009	.012		
Likelihood Ratio	15.186	3	.002	.012		
Fisher's Exact Test	10.989			.012		
Linear-by-Linear Association	2.533(b)	1	.111	.164	.082	.047
N of Valid Cases	20					

$P \leq 0.05$

Table 4.37 shows that the significant value is 0.009 which indicates that the frequencies of the four groups are significantly different at the level of 0.05. It can be interpreted that the test takers' opinions on the test scoring were statistically different.

Table 4.38 presents the frequency and percentage of the test takers' opinions on scoring.

Table 4.38: Frequency and Percentage of Test Takers' Opinions on the Scoring

Investigated topics	Frequency and Percentage	Groups			
		ACOM	ICOM	ACON	ICON
Scoring					
(positive)	Frequency	4	5	0	3
	Expected frequency	3	3	3	3
	Percentage	80	100	0	60
(neutral)	Frequency	1	0	0	2
	Expected frequency	2	2	2	2
	Percentage	20	0	0	40
(negative)	Frequency	0	0	5	0
	Expected frequency	0.3	0.3	0.3	0.3
	Percentage	0	0	100	0

N=5 for each group

Table 4.38 shows that the majority of Group ACOM (f=4), ICOM (f=5) and ICON (f=3) positively agrees on the test scoring while 100 per cent of Group ACON have a negative attitude towards scoring. Following are their responses.

The frequencies from the four group are presented in the following bar chart.

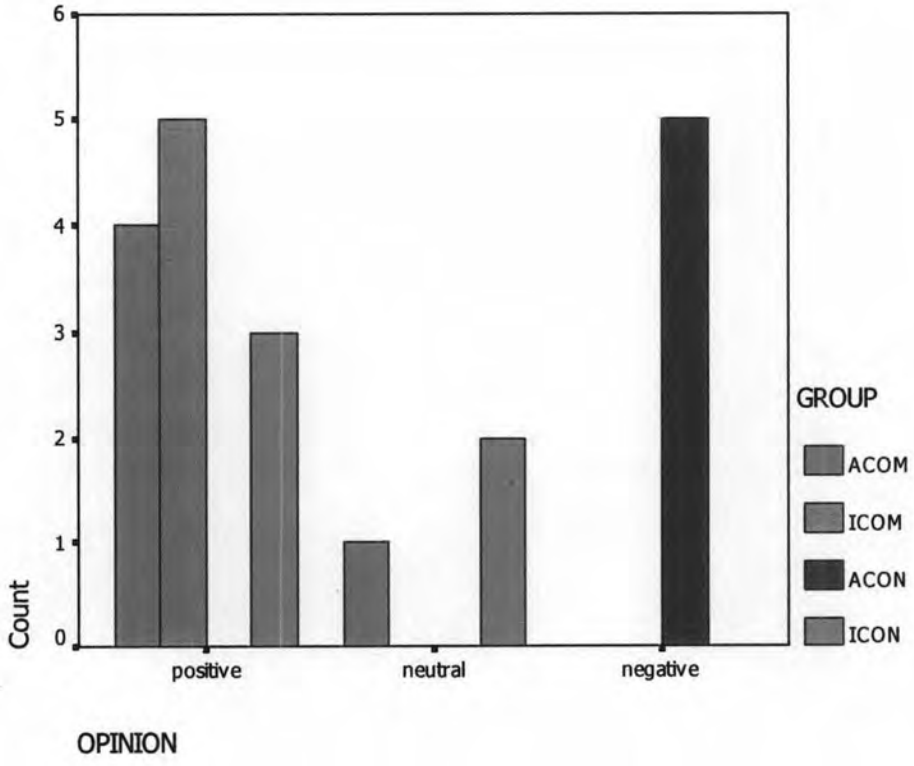

Figure 4.5: Difference frequencies among the four groups in terms of the test takers' opinions on the scoring

Qualitative responses: Group ACOM, ICOM and ICON reported the positive statements on the test scoring. They are summarized as follows:

The test scoring is effective because of using the computer. There might be some mistakes which occur if teachers mark the test by hand.

The criteria set in the program are standard so it is fair to all students.

Since the items are multiple choice, there is only one correct answer to each item. Therefore, there is no bias.

Qualitative responses: However, for the ACON group, all group members selected negative opinions. Their opinions can be summarized as follows:

I think scoring on multiple choice tests must be more effective because there might be more than one correct answer in each short answer item.

Teachers may spend longer time to score the tests. Moreover, students' scores are up to teachers' judgment. So, bias can occur.

3.3 *Interactivenes*

The students' opinions on interactiveness are given below.

Question: How does the test allow you to interact with the test tasks?

Table 4.39 presents the frequency and percentage of the test takers' opinions on the scoring.

Table 4.39: Chi-Square Tests of the Test Takers' Opinions on Interactiveness

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	12.724(a)	6	.048	.034		
Likelihood Ratio	16.980	6	.009	.017		
Fisher's Exact Test	11.636			.018		
Linear-by-Linear Association	.743(b)	1	.389	.483	.242	.080
N of Valid Cases	20					

$P \leq 0.05$

Table 4.39 shows that the significant value is 0.048 which indicates that the frequencies of the four groups are significantly different at the level of 0.05. It can be interpreted that the test takers' opinions on interactiveness were significantly different.

Table 4.40 shows the frequency and percentage of the test takers' opinions on interactiveness.

Table 4.40: Frequency and Percentage of the Test Takers' Opinions on Interactiveness

Investigated topics	Frequency and Percentage	Groups			
		ACOM	ICOM	ACON	ICON
Interactiveness					
(positive)	Frequency	1	0	3	3
	Expected frequency	1.8	1.8	1.8	1.8
	Percentage	20	0	60	60
(neutral)	Frequency	3	5	2	0
	Expected frequency	2.5	2.5	2.5	2.5
	Percentage	60	100	40	0
(negative)	Frequency	1	0	0	2
	Expected frequency	0.8	0.8	0.8	0.8
	Percentage	20	0	0	40

N=5 for each group

Table 4.40 shows that the majority of Group ACON (f=3) and ICON (f=3) positively agreed that the tests are interactive. The majority of ACOM (f=3) and ICOM (f=5) neutrally agreed on this issue.

The bar chart presents the differences among the four groups.

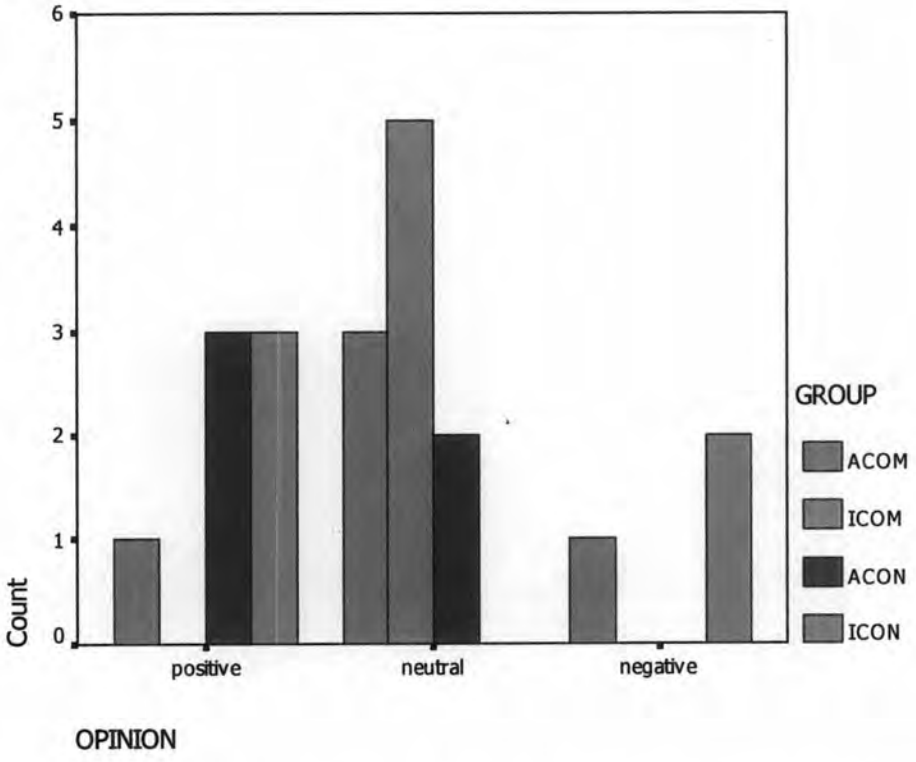

Figure 4.6: Difference frequencies among the four groups in terms of the test takers' opinions on interactiveness

Qualitative responses for Group ACON and ICON: They selected positive statements about interactiveness. Their responses are summarized as follows:

ICON: I felt as if I was involved in the test.

ACON: It is like we were listening to two people talking. I have to follow the conversation and think how I could answer them or communicate with them.

Qualitative responses for Group ACOM and ICOM: They reported neutral opinions on this issue. The statements are summarized as follows:

ICOM: I felt interactive to the test in some parts but not all.

ACOM: I think I am moderately interactive with the test.

3.4 *Authenticity*

When asked about authenticity, the students gave their opinions by selecting from statements as follows:

Question: How authentic are the test situations provided in the test? Are they relevant to your real life activities?

Table 4.41 shows Chi-Square tests of the test takers' opinions on the test authenticity.

Table 4.41: Chi-Square Tests of the Test Takers' Opinions on the Test Authenticity

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	16.000(a)	6	.014	.006		
Likelihood Ratio	17.454	6	.008	.011		
Fisher's Exact Test	11.864			.015		
Linear-by-Linear Association	5.677(b)	1	.017	.019	.010	.005
N of Valid Cases	20					

$P \leq 0.05$

Table 4.41 shows that the significant value is 0.014 which indicates that the frequencies of the four groups are significantly different at the level of 0.05. It can be interpreted that the test takers' perceptions on test authenticity are statistically different.

The following table illustrates the frequency and percentage of the test takers' opinions on the test authenticity.

Table 4.42: Frequency and Percentage of the Test Takers' Opinions on the Test Authenticity

Investigated topics	Frequency and Percentage	Groups			
		ACOM	ICOM	ACON	ICON
Authenticity					
(positive)	Frequency	5	3	3	1
	Expected frequency	3.0	3.0	3.0	3.0
	Percentage	100	60	60	20
(neutral)	Frequency	0	0	2	0
	Expected frequency	0.5	0.5	0.5	0.5
	Percentage	0	0	40	0
(negative)	Frequency	0	2	0	4
	Expected frequency	1.5	1.5	1.5	1.5
	Percentage	0	40	0	80

N=5 for each group

Table 4.42 shows that the majority of Group ACOM (f=5), ICOM (f=3) and ACON (f=3) positively agreed that the tests are authentic. Conversely, the majority of Group ICON gave negative opinions on this issue.

The following chart illustrates the differences compared among the four groups.

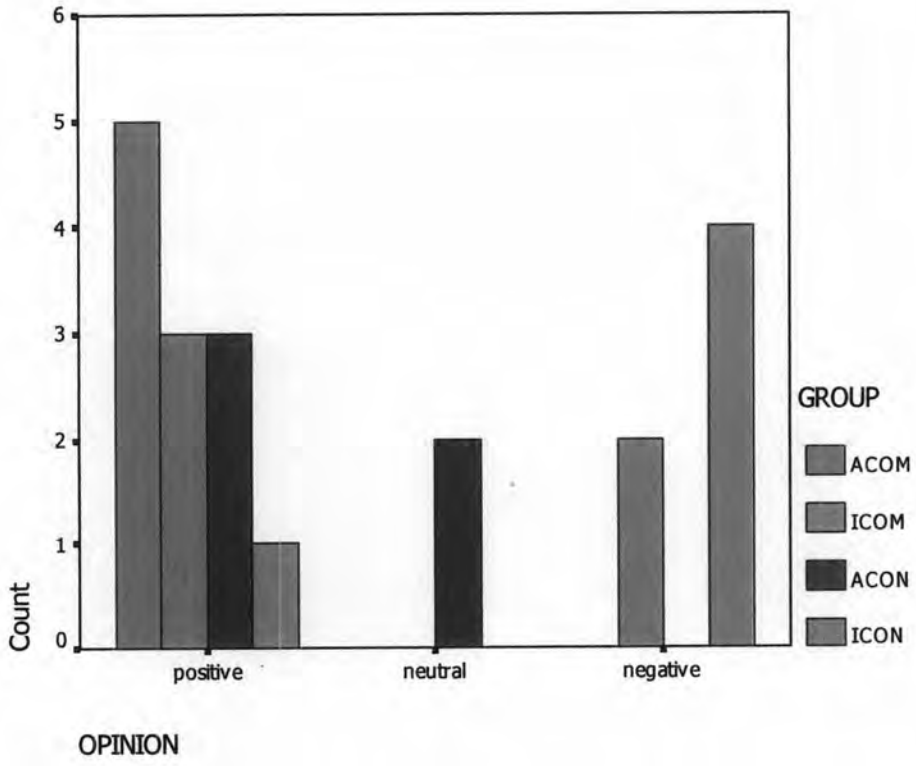

Figure 4.7: Difference frequencies among the four groups in terms of the test takers' opinions on the test authenticity

Qualitative responses: Group ACOM, ICOM and ACON selected positive statements on test authenticity. These are summarized as follows:

ACOM and ACON: The situation in the test is similar to our real life situations. We talk to friends or others about news or the things we read.

ICOM: The reading texts and job application are relevant to our real life.

Qualitative responses: However, the majority of Group ICON reported the negative opinions. Their opinions are given below:

ICON: It is just like an ordinary test. The test provides the texts and questions with choices. The test tasks are not authentic in our daily life.

4. Candidates' performance

This part deals with the students' attitudes towards the familiarity with the test, perseverance and overall attitude.

4.1 *Familiarity*

Question: Are you familiar with the test delivered by the computer?

Table 4.43 shows Chi-Square tests of the test takers' opinions on their familiarity to the tests.

Table 4.43: Chi-Square Tests of the Test Takers' Opinions on Their Familiarity to the Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	9.467(a)	6	.149	.151		
Likelihood Ratio	10.723	6	.097	.195		
Fisher's Exact Test	7.611			.195		
Linear-by-Linear Association	4.714(b)	1	.030	.036	.018	.009
N of Valid Cases	20					

$P \leq 0.05$

Table 4.43 shows that the significant value is 0.149 which indicates that the frequencies of the four groups are not significantly different at the level of 0.05. It can be interpreted that the test takers' opinions on their familiarity to the test were not statistically different.

Table 4.44 illustrates the frequency and percentage of the test takers' opinions on their familiarity to the tests.

Table 4.44: Frequency and Percentage of the Test Takers' Opinions on Their Familiarity to the Tests

Investigated topics	**Frequency and Percentage**	**Groups**			
		ACOM	**ICOM**	**ACON**	**ICON**
Familiarity					
(positive)	Frequency	2	2	3	5
	Expected frequency	3	3	3	3
	Percentage	40	40	60	100
(neutral)	Frequency	0	2	1	0
	Expected frequency	0.8	0.8	0.8	0.8
	Percentage	0	40	20	0
(negative)	Frequency	3	1	1	0
	Expected frequency	1.3	1.3	1.3	1.3
	Percentage	60	20	20	0

N=5 for each group

Table 4.44 shows that the majority of Group ACON (f=3) and ICON (f=5) positively agreed on familiarity with the tests. Equal numbers of Group ICOM members reported feeling both positive (f=2) and neutral (f=2) on familiarity. The majority of Group ACOM (f=3) reported negative opinions on this issue.

Qualitative responses: According to the chi square test, the frequencies among these four groups are not statistically different. The frequencies in Table 44 show the similarity of the frequencies in positive opinions. The positive statements are summarized as follows:

ACOM: When I studied English in the English Lab at my high school, I practiced doing the exercises that were similar to this kind of test.

ICOM: I took this kind of test when I studied in high school. The teachers had us study by computer-assisted instruction and take a test on the computer.

ACON: I feel familiar to the short answer items since I studied at the vocational diploma level.

Most of the tests I have taken since I finished high school are short answer. It is quite rare to see multiple choice tests, especially when I study here.

ICON: I had this multiple choice test since I was young. I can say I take this kind of test for my whole life.

4.2 *Perseverance*

As regards perseverance, the students gave their opinions reported below.

Question: How hard have you tried to do the test?

Table 4.45 shows Chi-Square tests of the test takers' opinion on perseverance.

Table 4.45. Chi-Square Tests of the Test Takers' Opinion on Perseverance

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	2.500(a)	3	.475	.871		
Likelihood Ratio	3.278	3	.351	.871		
Fisher's Exact Test	2.471			.871		
Linear-by-Linear Association	.238(b)	1	.626	.809	.405	.165
N of Valid Cases	20					

$P \leq 0.05$

Table 4.45 shows that the significant value is 0.475 which indicates that the frequencies of the four groups are not significantly different at the level of 0.05. It can be interpreted that the test takers indicated indifferently to the question about having perserverance on the tests

Table 4.46 presents the frequency and percentage of test takers' opinion on perseverance.

Table 4.46: Frequency and Percentage of Test Takers' Opinion on Perseverance

Investigated topics	Frequency and Percentage	Groups			
		ACOM	ICOM	ACON	ICON
Perseverance					
(positive)	Frequency	4	4	5	3
	Expected frequency	4	4	4	4
	Percentage	80	80	100	60
(neutral)	Frequency	1	1	0	2
	Expected frequency	1	1	1	1
	Percentage	20	20	0	40
(negative)	Frequency	0	0	0	0
	Expected frequency	0.3	0.3	0.3	0.3
	Percentage	0	0	0	0

N=5 for each group

Table 4.46 shows that the majority of Group ACOM (f=4), ICOM (f=4), ACON (f=5) and ICON (f=3) positively agreed with their perseverance in taking the tests.

Qualitative responses: Group ACOM, ICOM, ACON and ICON reported similar positive statements on perseverance. They are summarized as follows:

I tried hard to do this test. I thought that with my perseverance I could have a good result.

4.3 *Attitudes*

When asked about their attitudes towards the test, the students responded as follows.

Question: Generally speaking, do you have positive or negative attitudes towards this test?

Table 4.47 presents Chi-Square tests of the test takers' attitudes towards the tests.

Table 4.47: Chi-Square Tests of the Test Takers' Attitudes towards the Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	4.500(a)	6	.609	.871		
Likelihood Ratio	6.051	6	.418	.871		
Fisher's Exact Test	4.699			.871		
Linear-by-Linear Association	.000(b)	1	1.000	1.000	.561	.122
N of Valid Cases	20					

$P \leq 0.05$

Table 4.47 shows that the significant value is 0.609 which indicates that the frequencies of the four groups are not significantly different at the level of 0.05. It can be concluded that the test takers' attitudes towards the tests are not significantly different.

Table 4.48 shows the frequency and percentage of the test takers' attitudes towards the tests.

Table 4.48: Frequency and Percentage of the Test Takers' Attitudes towards the Tests

Investigated topics	**Frequency and Percentage**	**Groups**			
		ACOM	**ICOM**	**ACON**	**ICON**
Attitude					
(positive)	Frequency	4	5	3	4
	Expected frequency	4	4	4	4
	Percentage	80	100	60	80
(neutral)	Frequency	0	0	1	1
	Expected frequency	0.5	0.5	0.5	0.5
	Percentage	0	0	20	20
(negative)	Frequency	1	0	1	0
	Expected frequency	0.5	0.5	0.5	0.5
	Percentage	20	0	20	0

N=5 for each group

Table 4.48 shows that the majority of Group ACOM (f=4), ICOM (f=5), ACON (f=3) and ICON (f=4) has positive attitudes towards the tests. Their responses are listed below.

Qualitative responses: The majority of four groups has positive attitudes towards the tests. Their responses are summarized as follows:

ACOM: I have positive attitudes towards the test because I perceive that the scoring method is standardized and there is no human bias. The test directions are clear, so we know what we would do in the test without asking for any assistance. The test contents are interesting and relevant to my real life.

ICOM: I am positive towards this test. I feel fun taking this test. The results are immediately reported. The software also reports our weaknesses and strengths in reading ability.

ACON: I have positive attitudes towards this test because the situations in the test make me feel fun when I take it. I feel as if I were that character and had to think carefully to write an answer. The test also makes me realize about my writing skill that I should improve.

ICON: I feel positive towards this test. It is easy to do. Both those who understand and those who don't understand the texts can answer all test items. The questions are clear and not complicated.

In conclusion, the topics reported significantly different among the four groups are presented in the following table.

Table 4.49: Investigated Topics Reported as Significantly Different Among the Four Groups

Investigated topics	**Chi Square**	**Sig**
Test Fairness: the situations in the tests	13.429	.037
Time length	15.000	.002
Test contents	10.588	.014
The scoring	11.667	.009
Interactiveness	12.724	.048
Authenticity	16.000	.014

$P \leq 0.05$

4.3 Summary of the Results

The focus of this part was to report the results of data analysis. The data obtained from the four versions of English reading comprehension tests: ACOM, ICOM, ACON and ICON were analyzed for main effects and interaction effects using 2*2 ANOVA. From the data analysis, the crucial results obtained revealed that there was a significant difference found between the authentic tests (ACOM and ACON) and the inauthentic tests (ICOM and ICON). No significant difference was found between the tests delivered by computer (ACOM and ICOM) and the tests delivered by paper (ACON and ICON). However, the interaction effect between the test delivery mediums and the test authenticity was found. In terms of the effect sizes, Partial Eta Squared revealed that the test authenticity has a medium effect size and the test delivery mediums have a small effect size on the reading ability score.

Based on the opinions from the interviews, all students from the four groups expressed similar positive opinions on 1) the opportunity to demonstrate strengths and weaknesses in reading ability 2) perception of test fairness: appropriateness of the question 3) perception of nervousness while taking the tests 4) accuracy of the test to elicit their true ability in reading 5) accuracy of the test to elicit true ability in reading: the relevance of the test tasks to their real life reading activities, 6) test administration taken by students 7) test administration compared with other versions of the tests 8) facilities 9) familiarity 10) perseverance and 11) attitudes. All four groups also had similar neutral opinions on test difficulty. The negative opinions the four groups agreed with were on perception of test fairness: the effects on how to answer the test. Finally, there were the topics having significantly different and significant opinions on 1) test fairness: the situations in the tests 2) time length 3) test contents 4) scoring 5) interactiveness and 6) authenticity.