CHAPTER IV

FINDINGS AND DISCUSSIONS

This chapter discusses about the research findings and discussions. It provides the analysis and interpretation of data that had been collected to answer the research question about the consistency analysis of English teachers of Al Amin Islamic Boarding Senior High School Mojokerto in scoring essay test.

As explained in chapter III, grade eleven teachers held a weekly writing examination; an exposition essay test to know students' achievement in writing that kind of text. The students wrote the essay in handwriting to keep the originality of their writing. Even it might increases the subjectivity but the researcher had kept it by omitting the students' name before copying and giving to the examiners or raters; mean the teachers. Each six teachers or raters were asked to grade 10 papers of 40 essays which were chosen randomly by the researcher in pre- and post-scoring. 5 of 10 papers in post-scoring were the same essay that they had actually rated in pre-scoring. After two months interval, some teachers admitted that they did not remember about ever seen those papers before. In addition, the others said that although they remembered having ever seen the papers but they could not remember the grades that they gave.

After getting all teachers' pre-scoring in 22nd of February 2016 and post-scoring in 18th of April 2016, the researcher analyzed the data by using Cronbach

Alpha coefficient as descriptive statistic analysis and paired *t* test as inferential statistic analysis to check the significance of the finding and the null hypothesis test in SPSS. The study found varied result for each rater.

A. Findings

1. Descriptive Statistics

The raters were asked to grade five same essays in pre- and post-scoring with two months interval and some of them did not know that they score the same essay. The others might know that but they totally forgot what score they gave to each essay. Here is the table of five essays in pre- and post-scoring based on the rubric used in the assessment.

1.1.Table of Raters' Pre- and Post-Score of Five Same Essays

	No. Essay	Content		Organization		Grammar		Vocabulary		Mechanic		Total	
		PRE	POST	PRE	POST	PRE	POST	PRE	POST	PRE	POST	PRE	POST
RATER 1	1st	14	10	14	10	40	35	13	10	4	3	85	68
	2nd	11	11	10	10	32	30	10	11	3	3	66	65
	3rd	14	14	14	14	45	44	14	12	3	4	90	88
	4th	13	22	13	12	40	40	13	12	3	3	82	79
	5th	13	22	12	11	32	35	11	12	4	3	72	73
	1st	12.5	10.5	12	10	41	33	12.5	10	4.6	4.2	82.6	67.7

RATER 2	**2nd**	10.5	10	9.5	10	35	30	11.5	10	4.3	4	70.8	64
	3rd	13	13	14	13	46	41	14	14	4.3	4	91.3	85
	4th	14	13	14	13	43	35	13	10	4.6	3.5	88.6	74.5
	5th	12	13	11	13	38	41	12	12	4.6	4	77.6	83
RATER 3	**1st**	15	14	15	13	41	46	15	14	4.6	4.6	90.6	91.6
	2nd	14	12.5	14	12.5	38	40	14	14.5	3.8	4	83.8	83.5
	3rd	14	14	12.5	13	40	43	14	13	4.2	4.2	84.7	87.2
	4th	15	14	15	13	40	43	12.5	13	4.2	4.3	86.7	87.3
	5th	14	14	12	14	40	40	15	13	5	5	86	86
RATER 4	**1st**	12	11.5	12	11	38	35	11	11	4	3.8	77	72.3
	2nd	11	11	11	11	35	35	10	11	4	3.7	71	71.7
	3rd	11	11	11	11	36	35	11	11	3.7	3.7	72.7	71.7
	4th	12	12	12	12	38	40	11.5	12.5	3.8	3.8	77.3	80.3
	5th	11	11	11	11	35	35	10	11	3.7	3.7	70.7	71.7
RATER 5	**1st**	13	12	13	12	40	40	13	12	4	4	83	80
	2nd	10	14	13	12.5	40	36	13	12	4	4	80	78.5
	3rd	13	12.5	13	12.5	35	41	12	12.5	4	4.2	77	82.7
	4th	13	12.5	12.5	12.5	35	41	13	12.5	4.2	4	77.7	82.5
	5th	10	14	13	12.5	46	36	13	12	4	4	86	78.5
	1st	11	12.5	10.5	12	36	43	11.5	13	3.8	4.2	72.8	84.7

RATER 6	**2nd**	13	11.5	12.5	11.5	41	38	12.5	10	4.3	3.8	83.3	74.8
	3rd	13	11.5	12.5	11.5	41	36	12.5	10.5	4.3	4	83.3	73.5
	4th	12.5	13	11.5	12.5	40	43	11	12.5	3.8	4.3	78.8	85.3
	5th	13	11.5	12.5	11.5	41	36	12.5	10.5	4.3	4	83.3	73.5

The table above shows the real score of English teachers' pre- and post-scoring. The score of each category has agreed with the rubric given. Raters had meaning that the teachers who graded the essay. It was not mentioned and explained the teachers' identity in detail. The important one was they had same criteria; they had gotten a degree or language certificate. In other words, the teachers were admitted having the equal capability in English. Then, the meaning of essay number was the essay identity. Even the essays were given the number randomly in pre- and post-scoring, the five same essays had been put specific sign to help the researcher in analyzing them. Therefore, it was assured that those five essays in post-scoring were the same in pre-scoring.

There were many marks there. The way to read the table was horizontal, means from left to right. For example, the first column was 1st rater, the essay number column was 4th and the content column was pre = 13 and post = 22. It means that the first rater graded the content for 4th essay in pre-scoring was 13 but in post scoring was 22. Simply, the pre- and post-scoring produced different result, and so on.

By the table above, the researcher wanted to show that each rater has already changed in almost categories of the assessment. Post-score could be higher or lower than the pre-score. There were only a few post-scores which gotten as same as the pre-score. Absolutely, it was influenced the total score of each essay. The researcher could assure that the change was not unconsciously as some of them admitted that they forgot about the score given indeed the essay.

After collecting the data, the essay scores in the table above were analyzed the intra-class using SPSS 23 in Cronbach alpha coefficient analysis. Cronbach alpha coefficient of the teachers' consistency intra-rater reliability of five papers is presented below.

4.2 Table of Cronbach Alpha Coefficient Result of Raters' Intra-Rater Reliability Consistency

	Content	**Organization**	**Grammar**	**Vocabulary**	**Mechanic**	**Total**
RATER 1	0.426	0.682	0.925	0.158	- 1.333	0.832
RATER 2	0.813	0.768	0.684	0.696	- 0.468	0.722
RATER 3	0.571	-1.396	0.718	0.390	0.989	0.917
RATER 4	0.947	0.750	0.708	0.800	0.375	0.796
RATER 5	-10.435	- 0.667	- 4.850	- 2.400	- 0.667	- 3.969
RATER 6	- 3.000	- 2.444	- 3.800	- 6.087	- 10.556	-10.248

The table showed the statistical analysis of data in the previous table. The result was still formed in numerical. All five essays scores in previous table were analyzed in SPSS 23 based on the categories. For instance, the 2nd

rater's grammar scores of five essays were put and analyzed using SPSS software and produced 0.684 as the result of Cronbach alpha coefficient. As the various score, each rater got different result in all categories. Unfortunately, some data produced negative value and absolutely it will be rounded into 0.00 score. Nevertheless, the researcher used this real result to analyze whether the raters were consistent or not.

To know the intra-rater reliability level of each teacher, the result of Cronbach alpha coefficient formed in numerical data analysis was interpreted based on the Reliability Interpretation presented in Data Analysis Technique subsection. The interpretation result is presented below.

4.3 Table of Raters' Reliability Interpretation Result

	Content	Organization	Grammar	Vocabulary	Mechanic	Total
RATER 1	Enough	High	Very High	Very Low	Unreliable	Very High
RATER 2	Very High	High	High	High	Unreliable	High
RATER 3	Enough	Unreliable	High	Low	Very High	Very High
RATER 4	Very High	High	High	High	Low	High
RATER 5	Unreliable	Unreliable	Unreliable	Unreliable	Unreliable	Unreliable
RATER 6	Unreliable	Unreliable	Unreliable	Unreliable	Unreliable	Unreliable

The numerical result of Cronbach alpha coefficient was interpreted in the table above. The interpretation was made by checking the numerical result with the range score in the table 3.1. For example, 3rd rater got .917 in Cronbach alpha coefficient; means that he got Very High level of reliability as

the score was included in the range of $0.800 < \alpha \leq 1.000$. Therefore, the table above presented each result in the word.

As Table above shows, 1st rater got Very High reliability level in Grammar whereas High level in Organization. Besides, it indicated reliable Enough for Content but Very Low in Vocabulary. Mechanic was the worst as it got negative score which means that it was very unreliable. Luckily, his total score was very reliable as it got Very High level.

2nd rater was different from the 1st. Most of categories got High level, such as Organization, Grammar, Vocabulary and it might influence the Total score. The most reliable was in Content because 0.831 means it existed in "Very High" level. In contrary, Mechanic presented unreliable as it got negative score like the 1st rater.

Mechanic and the Total score of 3rd rater were almost perfect as it presented Very High reliability. He got High level in Grammar, reliable Enough in Content and Low level in Vocabulary. Unfortunately, it was the same as two raters before that they have minus value in one of their categories which means unreliable, this rater was in Organization.

This rater was the most stable than the others, 4th rater. It did not have negative value for any categories. Four categories had achieved High level, like Organization, Grammar, Vocabulary and absolutely it were impacted to the Total. The highest level, almost perfect, was Content. Besides, Mechanic

was the lowest reliability level as it got .375 score and it was still in positive value.

The worst unreliable raters are 5th and 6th raters. All of their categories presented negative value. It can be said that they got 0.000 score or were admitted as zero reliability. Simply, it was regarded that they were included in inconsistent or unreliable level.

As the various marks gotten, it was needed to make the average of all grades so that it could conclude the result which represented and covered all raters in all categories. Here is the table of average result. The table shows the average result of pre- and post-scoring of five same essays. For example, the pre-scoring of 1st rater was the average result from all pre-scoring in all categories and so was the post-scoring.

4.4 Table of All Essays' Average Results

No. Essay	PRE	POST
1st	81.8	77.4
2nd	75.8	72.9
3rd	83	80.7
4th	81.9	81.5
5th	79.3	77.6

After getting the average, it would be examined the Cronbach alpha coefficient for the intraclass correlation. It was aimed to know the overall result means the conclusion result of all raters. The way of calculating the coefficient was as same as previous computation. The result showed that it was very consistent as it got .942 in both Average Measure of intra-class correlation and Cronbach alpha. The value was more than 0.7 which means that English teachers or raters had good reliability consistency. Even there are two raters got inconsistent in all categories but it did not influence other results that affect the average.

4.5 Output of Intraclass Correlation in Cronbach Alpha Coefficient of SPSS 23

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.942	.948	2

Intraclass Correlation Coefficient

	Intraclass Correlation[b]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	,890[a]	.283	.988	17.186	4	4	.009
Average Measures	**,942[c]**	.441	.994	17.186	4	4	.009

2. Inferential Statistics

After getting the result of descriptive statistics, the finding would be checked the significance by using inferential statistics in paired *t*-test of SPSS 23. The value checked was not the whole result but it was only the average result as it has covered all values of all raters in pre- and post-scoring. This is the result of paired *t* test in SPSS 23.

4.6 Output of Paired T Test Result Of Average Result in SPSS 23

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	PRE	80.360	5	2.8867	1.2910
	POST	78.020	5	3.3937	1.5177

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	PRE & POST	5	.902	.036

Paired Samples Test

		Paired Differences					T	Df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	PRE – POST	2.3400	1.4775	.6608	.5054	4.1746	**3.541**	4	**.024**

The first table is Paired Sample Statistics that showed the statistic summary of pre- and post-scoring. The table provides that the average score in pre-scoring was 80.360 and in post-scoring was 78.020. It indicated reduction for about 2.340. The standard deviation presented the data variation in each variable, that in pre-scoring was 2.887 and in post-scoring was 3.394. Also N was the number of data which there were five essays graded twice by raters in two-week interval.

Paired Sample Correlation showed the correlation between two variables that produce 0.902 with 0.036 for the significance. It means that the correlation between pre- and post-scoring was so related.

The last is Paired Sample Test. It can be interpreted as:

- Hypothesis

  H_0 = the intra-rater reliability of English teachers at Al-Amin Islamic Boarding School Mojokerto in scoring essay test is not consistent.

  H_1 = the intra-rater reliability of English teachers at Al-Amin Islamic Boarding School Mojokerto in scoring essay test is consistent.

- Significance level

  Sig = 0.05

- Critical area

  Based on *t*-test:

  Reject H_0 = *t*-test > *t*-table (5%, N-1)

Accept H_0 = *t*-test < *t*-table (5%, N-1)

Based on p-value (Sig.):

Reject H_0 = *p*-value < 0.05

Accept H_0 = *p*-value > 0.05

- Decision

  t-test = 3.541 > *t*-table (5%, N-1) = 2.776;

  Sig. = 0.02 < 0.05;

  means H_0 is rejected.

  The intra-rater reliability of English teacher at Al-Amin Islamic Boarding School Mojokerto in scoring essay test was consistent.

B. Discussion

According to the finding, teachers were mostly very consistent in scoring the same paper. Table 4.3 showed that most of raters achieved upper consistency level beside only some raters got enough and lower level. The most consistent rater in all categories was the fourth. Even the fourth rater achieved Low level in Mechanic but no one categories got negative value or zero reliability that means inconsistent. The first rater was very consistent in Grammar and Total, but did not do well in his ratings of Mechanic. The second was almost the same as the fourth yet he was very inconsistent in Mechanic too, like the first. The third was very consistent in Mechanic and the Total score as both got Very High level but very

inconsistent in Organization. Unfortunately, the fifth and sixth teacher or rater seemed to be the least consistent. Their ratings were abysmal in all categories. In fact, they even contradicted in their own ratings in the pre-scoring so that the coefficient is negative.

In order to be easy in taking the conclusion, all various results were taken the average and calculated in Cronbach alpha coefficient. Based on the reliability interpretation, it produced Very High consistency as it got .924 of intraclass correlation in SPSS 23. This value was more than 0.7 as the standard of Cronbach alpha coefficient in deciding the reliability. Simply, it was proved that English teachers of Al-Amin Islamic Borading School Mojokerto had good reliability. In addition, paired t-test result as the significant calculation of inferential statistic also qualified the rules of rejecting the null hypothesis. The rules are: 1) t-test was more than t-table; 3.541 > 2.776 and 2) Sig. = 0.02 was less than 0.05 as the level of significant. It means that the result was the real score, not incidentally. Even there were two raters got inconsistent or unreliable in all categories but it did not give any impact to the calculation which proven that the intra-rater reliability of English teachers at Al-Amin Islamic Boarding School Mojokerto was internally consistent. It can be said that the inconsistent scores gotten happened by chance with many exceptions from the raters' self that can be investigated in the next research.

This result was very different from the intra-rater reliability result in the journal from Viphavee Vongpuvimitch entitled *Classroom Writing Teachers'*

Intra and Inter-rater Reliability: Does it matter? The journal result showed that the analytic raters got low intra-rater reliability which means that almost raters were not very consistent in all categories. It did not present the specific score that was defined the level of consistency as the consistency admission for this study was only when the rater got .800 – 1.00 score. This journal did not calculate the conclusion result of all raters' consistency so that there were not the real and specific numerical criteria to decide whether "they" were consistent or not. Besides, it did not hold an inferential statistic test so that it was questionable whether the result produced incidentally or not. In addition, the time interval between pre- and post-score was only one week when the risk of some carryover had big possibility occurred.

Another journal, *Rater Discrepancy in the Spanish University Entrance Examination* by Marian Amengual Pirazzo[1] from University of Balearic Island had same result as this study that the intra-class correlation mean was high. The difference was from the result of significance. It indicated that there were not significant differences between the holistic pre- and post-scoring. It might be happened because the time interval was too long; three months, and this was one of the risks of carryover. Therefore, even if the intra-class reliability was quite high but there were not significant differences among the result, means that it only happened by chance. Besides, the scoring tool of the essay was the opposite

[1]Marian Amengual Pirazzo. *Rater Discrepancy in the Spanish University Entrance Examination*. Journal of English Studies University of Balearic Island Vol.4 page 23-26, 2003-2004.

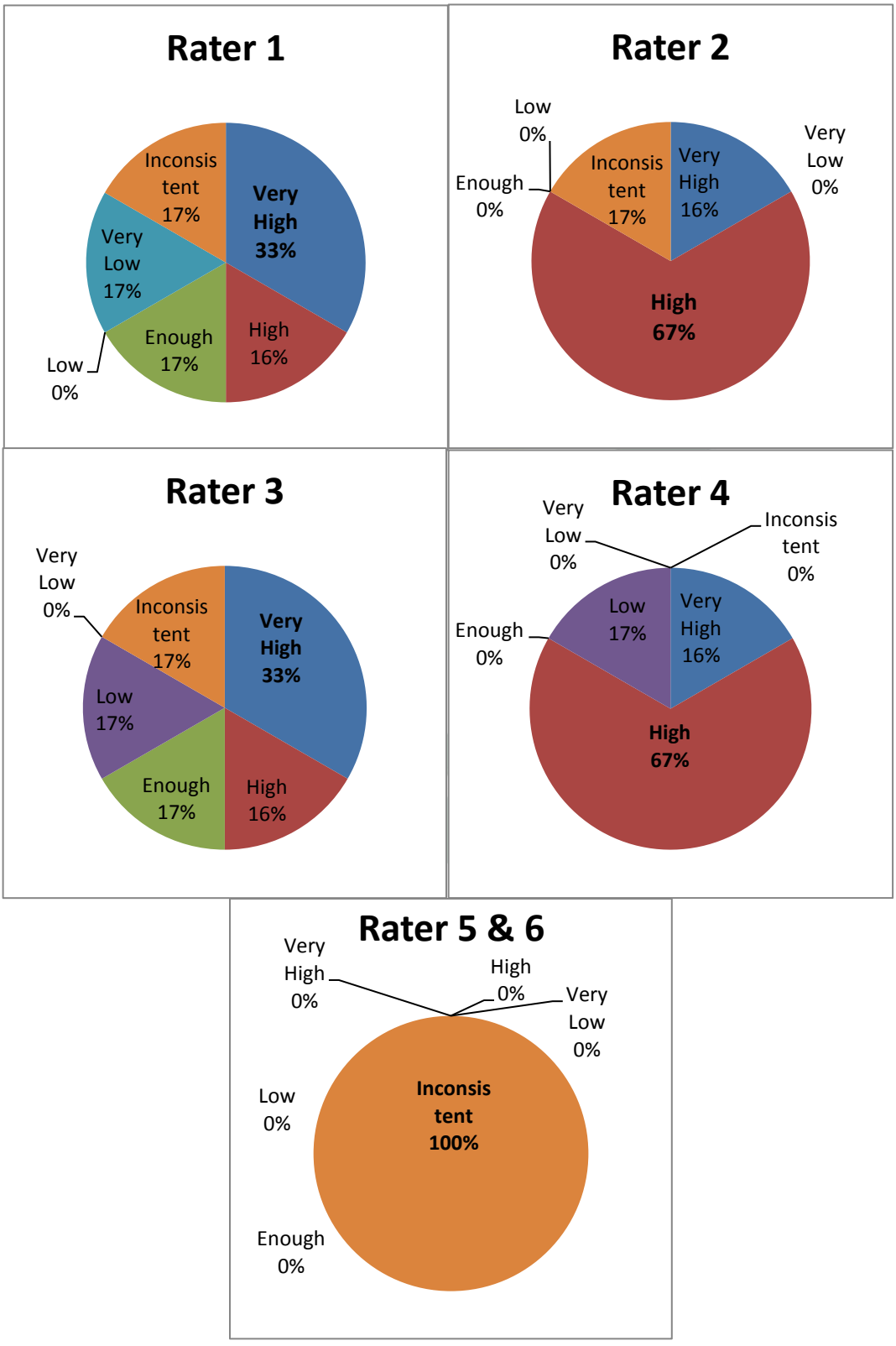
from this study. This journal used holistic whereas the study used analytic. Many factors that make the result of different scoring tool in two studies were different. Therefore, it was needed to investigate deeper in the next research whether the tool used influence the result or not.

Although this study used rubric for the assessment, it was different from journal of *Reliability and Validity of Rubrics for Assessment through Writing* by Ali Reza Rezaei from California State University and Michael Lovorn from The University of Alabama, USA[2] that showed the rubric impact in improving raters' reliability and validity. As the newest study about this topic in Indonesia, especially in Sunan Ampel State Islamic University, this research did not investigate it too far. The function of rubric was only to help teachers or raters having the same and specific criteria in assessing writing.

To give detail information about reliability level reached by each teacher, the study presented in diagrams what level they were achieved and its percentage.

[2] Ali Reza Rezaei – Michael Lovorn. "Assessing Writing". *Reliability and Validity of Rubrics for Assessment Through Writing* Vol. 15, 2010.

4.1 Circle Diagrams of Raters' Reliability Level Percentage

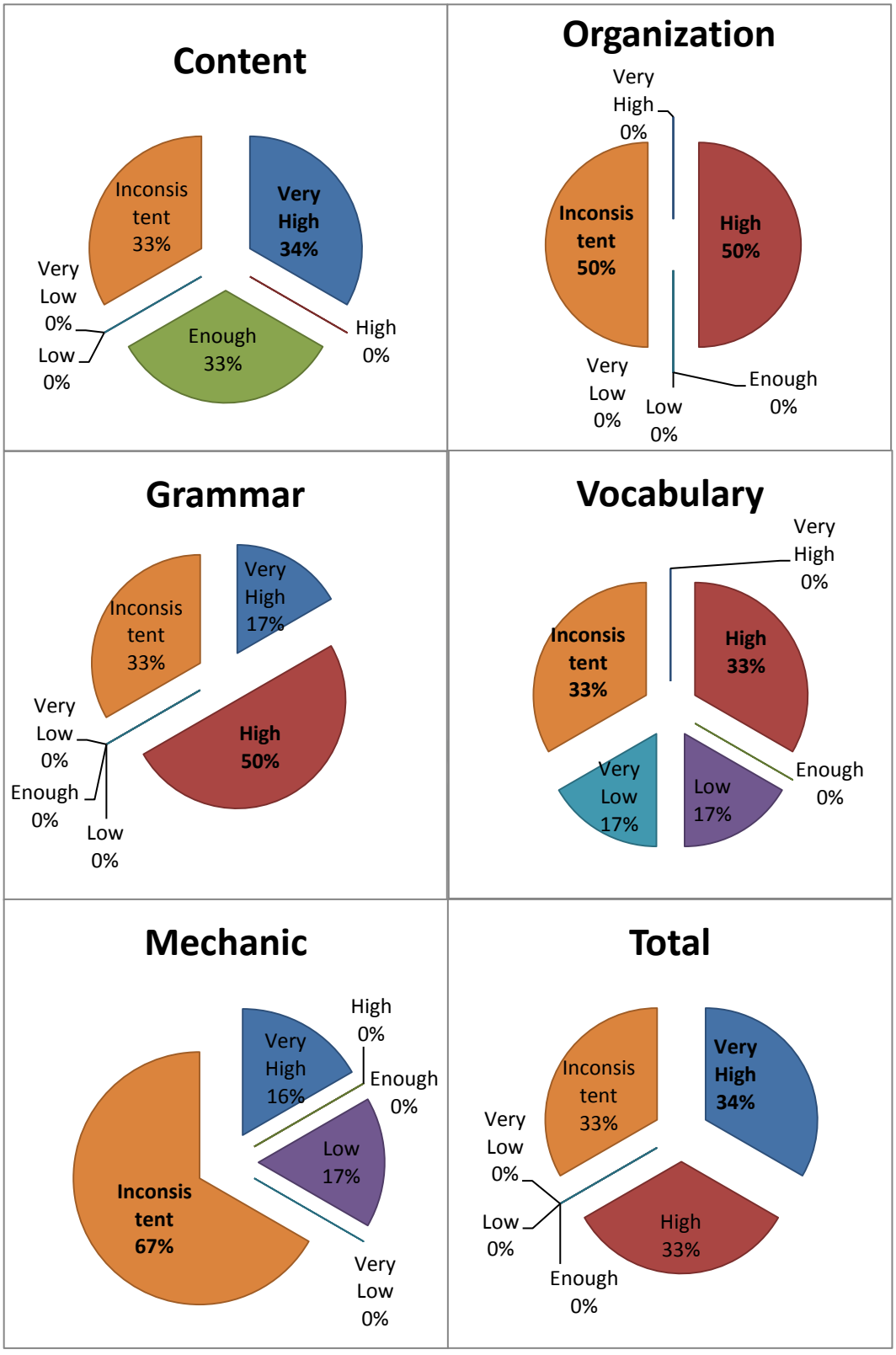


The diagrams above draws clearly that 1st and 3rd teacher got Very High level in 33% of all categories. In addition, 67% High level was achieved by 2nd and 4th rater. Those percents were the highest percentage earned by each rater. It was enough to prove that most of raters in almost category got upper consistency level so that the average result which represents the final conclusion produced very high value. Even if 1st, 2nd and 3rd teacher had reached good value in Very High and High level but they still got negative value in one of categories, especially Mechanic. Fortunately, the negative did not give big impact in the final result and it can be said that the minus score happened incidentally.

In addition, although 4th rater did not get any values in Very High level but she had no minus score in all categories who is admitted as the most stable rater of all. Besides that, 5th and 6th rater had proven well to be the most unreliable rater of all. As it has explained before that it did not influence the average result of this study, both raters might be disturbed by internal or external factors, such as bias, illness, etc., during scoring the essay that can be analyzed deeply in the next research.

As the various result reached by each rater, the researcher tried to map in diagrams what category that have to be given more attention in scoring essay test. It was aimed to know their weakness so that they can improve their reliability.

4.2 Circle Diagrams of Raters' Reliability Level for Each Category



Diagrams above told that Very High level was reached well in Content and the Total score whereas High level was achieved in Grammar only. It means that Content and Grammar were the most objective categories of all. The other categories got High level as the same as inconsistent level, like Vocabulary. It can be said that each rater had different views in this category that can cause significant different of raters' achievement. Even if half raters got inconsistent level in Organization but another half got High level here, unlike the Vocabulary that produce more varied result. Therefore, the worst categories were Mechanic as 67% of teachers got unreliable level (minus score) here. In other word, Mechanic was not considered well by English teachers in scoring essay test so that more than a half of them got zero reliability.

The teachers did a better job in grading Grammar that was influenced the Total. Almost raters reached Very High and High consistency level as the most objective of the five subscales. In contrast, for about a half raters were very consistent in Organization but the other half were inconsistent. Also even if two teachers reached Very High level in Content but the other two were only stuck in Enough indeed another were inconsistent. The least objective of all were Vocabulary and Mechanic as almost all raters achieved unsatisfactory consistency level.

Even if the diagrams told about the raters' reliability in each category but it did not compare their achievement in each category, means the inter-rater reliability. It just presented the percentage of reliability level in each category to

define what category must be given more attention to improve teachers' reliability so that it would not get lower consistency level.

According to all findings and discussions presented in this chapter, the result of this study showed that the intra-rater reliability of English teachers at Al-Amin Islamic Boarding School Mojokerto was consistent in scoring essay test. The negative result of some raters was admitted as the incident that was happened in chance. From here, it seems that it is very difficult to derive consistent result from raters. Even if they have many experiences in scoring essay test for many years, it is not assure that they have good reliability in scoring subjective test, like essay test.

The result of this study was tentative. The limitation of this study has explained in Chapter I that the result can be changed in another chance as the scoring process was only twice, pre- and post-scoring. However, it is wished that this study allow all language teachers, raters or testers; especially English language, to consider the importance of rater reliability or consistency in writing assessment, like essay evaluation marking. Also, this research is hoped that it can serve as an example for further research in the same topic to eliminate everything that can influence the objectivity of scoring, particularly scoring subjective test.