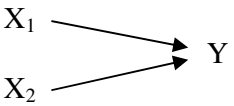
III. RESEARCH METHODS

This chapter discusses five core elements of methodology. They are (1) design of the research, (2) population and sample of the research, (3) research instrument, (4) research procedures, (5) criterion of a good test, (6) data treatment, and (7) hypotheses testing.

3.1. Design of the Research

In designing this study, the researcher adopted *one shot study design*. Based on the research question in this study, handwritten and computer test are independent variables; while raters' score is dependent variable.

Here is how the design looks like:



Where,

X : Medium of presentation

X1 : Handwriting

X2 : Computer typing

Y : Raters' scores

(Setiyadi, 2006 : 131)

3.2. Population and Sample of the Research

The population in this research were essays produced by one class of third semester students in English Study Program of Lampung University. There were forty one students in that class as the population. All of the students were following the training session and the writing task. At the end of the writing task, there were forty one essays produced by forty one students.

In order to choose the research sample, the researcher used random sampling by simply drawing lottery for twenty one (21) essays out of forty one essays (41). The 21 sample size was selected randomly to meet the exact number of sample in which the raters would be scoring.

3.3. Research Instrument

In collecting the data the writer used the following instrument:

Writing Task

In this writing task, the participants were asked to hand-write one page length of an argumentative essay from a given prompt. The writing task was conducted on September 16th, 2014 with 41 students participated. The participants were given 2x45 minutes time allocation to finish their essay based on the prepared prompt.

Prior to the writing task, the researcher delivered a training session to explain criteria of a good argumentative essay. The training session was done on September 9th 2014 in which the students learnt about components of argumentative essay. For further details, see the training session in **Appendix 2.**

The next meeting, which was the following week on September 16th 2014, the researcher administered the writing task based on the prompt prepared before. The prompt itself was created from the current material the students were discussing in the Literature class. The prompt was about an old epic from Anglo-Saxon era which tells about the ancient hero, Beowulf. For further details, see the prompt in **Appendix 3.**

3.4 Research Procedure

The procedures in this research were conducted in the following sets (1) determining the sample of the research, (2) administering training session, (3) administering writing task, (4) transcribing original essay, (5) distributing the original and transcribed essay, (6) scoring essays by raters, (7) analyzing the data, and (8) drawing findings and conclusions from the data.

1. Determining the Sample of the Research

   The sample would be taken from 21 essays out of 41 essays produced by students of English Study Program in Lampung University. The essays were made during Introduction to Literature class.

2. Administering Training Session

   The training session was done on September 9th, 2014 prior to the writing task. In this session the researcher delivered the material about components which make up a good essay.

3. Administering Writing Task

   The writing task was conducted on September 16th, 2014 in the same class. The students were asked to handwrite one full page essay on paper which is around 400 words. The instruction was made clear; the students composed the essay from the given prompt. The prompt was made to meet their understanding about the topic they had learned; it was under the topic of 'Beowulf', an epic poem from Anglo-Saxon era. The students were given 2x45 minutes time allocation to finish and submit their essays.

4. Transcribing Original Essay

   It was transcribing the original handwritten format into computer-text format. The transcribing was done verbatim (including all spelling, grammar, and punctuation errors) into computer format by the research team.

   To ensure the precision in transcription, the following procedures were adopted. When transcribing responses from their original handwritten form to computer text, responses were first transcribed verbatim into the computer. The transcriber then printed out the computer version and compared it word by word with the original, making corrections as needed. A second person then compared these corrected transcriptions with the originals and made additional changes as needed.

5. Distributing the original and transcribed essays

   After transcribing the original handwritten. the essays were distributed to six raters (3 pairs) following the format below:

Table 3.1 Essay Distribution Among 6 Raters

Essays	**Essay Format**		
	Handwritten	Single-spaced computer text point 12 TNR font	Double-spaced computer text point 14 TNR font
#1 - 7	Rater 1, 2	Rater 3,4	Rater 5,6
#8 - 14	Rater 5,6	Rater 1,2	Rater 3,4
#15 - 21	Rater 3,4	Rater 5,6	Rater 1,2

From the table of distribution above, the total sample size were 63 essays; 21 for the original handwritten, 21 for single space, and 21 for double space. We can also see that none of pairs or raters scored twice for the same essay. They were also unaware of the presentation effect as being the core of this investigation.

From the distribution table above, we can see that rater 1 and 2 scored the handwritten essay for essay number 1-7, computer text form single space 12 point for essay number 8-14, and computer text double space 14 point for essay number 15-21. Rater 3 and 4 scored the single spaced 12 point essay number 1-7, the double spaced 14 point essay number 8-14, and handwritten format for essay number 15-21. While the rest two raters, rater 5 and 6 scored the double spaced 14 point essay number 15-21, the handwritten form number 8-14, and single spaced 12 point essay number 1-7. In the end, the score gained from each pair would be regarded as one average total score.

6. Scoring Essays by Raters

   There were six raters employed in this research. Four of them were advanced graduate students in several state universities in Indonesia. The other two raters were English instructors in a Language Testing Center in a local university. After all, there were three pairs of raters working on different format of essay, but none of them scored the same essay twice.

7. Analyzing Data

   The researcher used one way ANOVA to analyze the data. The data were statistically computed through the Statistical Package for Social Science (SPSS) version 19.

8. Drawing findings and conclusions

   The last step of this research was drawing findings and conclusions from the data analysis above. In this step, the researcher also formulated some suggestions and recommendations for further research.

3.5 Criterion of a Good Test

In analyzing the data, the researcher used one way ANOVA to measure more than two or three groups of mean, they are raters' scores on: original handwritten, computer text single spaced point 12, and computer text double spaced point 14.

a. Validity

A good test can be seen from its validity. Validity refers to which an instrument really measures the objective and suitable with the criteria (Hatch and Farhady,

1982: 250). The validity of this research will be seen from content and construct validity.

Content Validity

In order to meet the content validity, the researcher applied rubric to assess the essay. The rubric was selected because it has been widely applied to assess the performance of English Foreign Learners (EFL) in the states.

Following a scoring procedure for composition items for ESL/EFL students, all responses in a given format were multiplied by seven for Idea Development and Organization criteria, and were multiplied by three for Grammar and Sentence Structure criteria. The scoring guidelines for the composition items focused on two areas of writing, namely Idea Development and Organization, and Grammar and Sentence Structure. Both scale for Idea Development and Organization and the scale for Grammar and Sentence Structure ranged from 0 to 10 and were multiplied by seven and three respectively. Table below presents the category descriptions for each point on the two scales.

Table 3.2: Specification on Data Collecting Instrument for EFL

Q1 - Development and Organization (multiply rating by 7 points)

			Rater's Comments:		
Overall Content --Did you cover all aspects of the prompt? --Did you use topic sentence(s) --Did you organize your answer well and/or use transitions?					
Excellent	Good	Average	Needs Improvement	Unacceptable	Score x7

10	9	8	7.5	7	6	5	4	3	2	1	0	

Q1 - Grammar and Sentence Structure (multiply rating by 3 points)

Are your sentences complete and correct?
Are your sentences clear and easy to understand?

Type of Mistake	How many times?	**Rater's Comments**
art=article use		
frag=fragment		
cap=capitalization		
pos=possessive		
prep=preposition		
pro=pronoun		
p=punctuation		
ros=run on sentence		
sva=subject/verb agree		
sn/pl=singular / plural		
sp=spelling		
vt=verb tense		
mod=modal use		
wf=word form		
inf/ger=infinitive/gerund		
wo=word order		
wc=word choice		
wm=word missing		
ss=sentence structure		

Overall level of interference with meaning:												
Excellent	Good		Average	Needs Improvement				Unacceptable				Score x3
10	9	8	7.5	7	6	5	4	3	2	1	0	

Source: Content-based writing rubric for EFL Students (Doc. of Missouri State University)

Construct Validity

Construct validity measures whether the construction had already referred to the theory and objectives or not (Hatch and Farhady, 1982: 251). The rubric presented above has met the concept of writing assessment as discussed in Chapter 3.

b. Reliability of the Raters/Inter-rater Reliability

Since this research employed multiple raters in assessing students' essays; thus the reliability of the raters is very important to measure. The reliability of raters is known as inter-rater reliability which is a measure used to examine the agreement between two people (raters/observers) on the assignment of categories of a categorical variable. It is an important measure in determining how well an implementation of some coding or measurement system works. In this ANOVA based research, the researcher used Kappa for assessing the reliability of agreement between a fixed number of raters when assigning categorical rating to a number of items or classifying items. The measure calculates the degree of agreement in classification over that which would be expected by chance.

Below is the formula of Cohen's Kappa Inter-rater Reliability Coefficient:

$$\boldsymbol{K = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}}$$

Where,

K : Kappa reliability coefficient

Pr(a) : Actual observed agreement,

Pr(e) : Chance agreement

In accordance with the content based rubric (**see Appendix 4),** there are four scales that raters needed to score the essay: 1-30 for unacceptable, 40-70 for needs improvement, 75 for average, 80-99 for good, and 100 for excellent.

Below is the description of the rating scale:

Table 3.3 Rating Scale Distribution

Value	Scale	Description
0	0 - 30	Unacceptable
1	40 - 70	Needs Improvement
2	75	Average
3	80 - 99	Good
4	100	Excellent

The data of ratings were then calculated using SPSS version 19 to find out the percent agreement among raters. The higher the percent agreement, the more reliable the raters are.

As shown, the table below presents the result of *Kappa Correlation Coefficient* which was statistically calculated with SPSS version 19.

Table 3.4 Frequency Table of Inter-Rater Agreement

The Rating Difference

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	-1	2	3.2	3.2	3.2
	0	57	90.5	90.5	93.7
	1	2	3.2	3.2	96.8
	2	2	3.2	3.2	100.0
	Total	63	100.0	100.0	

From the frequency table above we can see that 0 valid in the first column of the table means that raters put the same perception on rating scale. The frequency showed a high frequency, that was 57 out of N = 63 which suggested 90.5 % of

agreement. The variable that is not zero (0) is identified as the difference perception between raters somewhere in the scoring or it can be said there was disorder in the agreement. However, the number showed a low frequency which was only about 9 %.

Below is the table that shows how *Kappa* can analyze the percentage of *count* and *expected count* of agreement. Prior to the calculation, the researcher grouped the six raters into two pairs: rater A consists of rater 1, 3 and 5; rater B consists of rater 2, 4, and 6. The idea behind grouping these raters was because the valid percent in the previous table above showed a high percentage of validity, thus all raters are relatively comparable.

Table 3.5 Cross-tabulation Percent of Agreement

Rater A Rating * Rater B Rating Cross-tabulation

			Rater B Rating			
			Needs Improvement	Average	Good	Total
Rater A Rating	Needs Improvement	Count	**46**	1	0	47
		Expected Count	36.6	7.5	3.0	47.0
	Average	Count	1	**8**	1	10
		Expected Count	7.8	1.6	.6	10.0
	Good	Count	2	1	**3**	6
		Expected Count	4.7	1.0	.4	6.0
Total		Count	49	10	4	63
		Expected Count	49.0	10.0	4.0	63.0

The expected count means that the expected chance set by the null hypothesis. If in the count percent the number is higher than the expected count, it means that the raters agree above chance. From the cross-tabulation table above, we can see that the null hypothesis set 36.6% of agreement for category **Needs Improvement,** but from the count percent we can see that both raters agree 46% which means there was an improvement in agreement. We can also see that in **Average,** the expected count was 1.6% for average category, but both raters rated 6% of agreement, which also means the level of agreement for **Average** category was above chance. There was only 1% difference, that was when rater 2 rated as 1% as average, while rater 2 rated as needs improvement. When we look at in **Good** category, the expected count was .4%, while both raters seemed agree for 3% level of agreement for **Good** category. From the table above it can be concluded that the level of agreement was good because both raters agree beyond chance. The *Kappa* then estimated the above chance above as the value level which was presented in table below.

Table 3.6 Symmetric Measures

Symmetric Measures

	Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Measure of Agreement Kappa	.755	.089	7.663	.000
N of Valid Cases	63			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

From the symmetric measures above, it can be said that the inter-rater reliability was valid in the level of .755 with p value was less than .05.

3.6 Data Treatment

In running one way ANOVA, there are five data assumptions that should not violate in order to support the result of the ANOVA calculation (Setiadi 2006: 173). They are:

1. There is only one dependent variable and one independent variable with three or more level. In this research, the dependent variable is the raters' scores and the independent variable is the essay formats with three type of treatments, they are handwritten, single space point 12 TNR fonts, and double space point 14 TNR fonts. So, the first assumption is not violated.
2. The dependent variable should be measured at the interval/ratio level. In this study, the dependent variable is continuous variable, that is the scores awarded by raters, and it is ranged from 0-100. Therefore the second assumption is met.
3. It is a between group comparison. In this research the independent variables are the subjects to compare. So, the third assumption is not failed.
4. The dependent variable should be approximately normally distributed for each category of the independent variable. In this research, the researcher employed *Shapiro-Wilk test of normality* which is available on SPSS and because this type of normality test is the most appropriate one for a research with sample size less than 50; however it can handle sample sizes as large as 2000.

Table 3.7 Test of Normality

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Essay Format	Statistic	Df	Sig.	Statistic	df	Sig.
Rater's Score	Handwritten	.248	21	.002	.778	21	.000
	single space point 12	.241	21	.003	.802	21	.001
	double space point 14	.137	21	.200*	.911	21	.056

a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.

The table above explained the result of *Shapiro-Wilk test of normality.* We can see from the table that for the "handwritten", "single space point 12", and "double space point 14" format group, the dependent variable "raters' scores" was deviated. It was shown by the significance value of less than 0.05 (only one category met this assumption), so the data has non-normal distribution. Fortunately ANOVA only requiring approximately normal data because it is quite "robust" to violations of normality, meaning that assumption can be a little violated and still provide valid results.

5. The number of sample size is not too small (at least 5 data for each cell). In this research the sample data for each category is 21. So, the last assumption is not violated.

3.7 Hypothesis Testing

The hypothesis was statistically analyzed using One Way Anova that draws the conclusion in significant level if $P > 0.05$, H_0 accepted, and $P < 0.05$, H_1 accepted.

H_0 : There is no difference on raters' score for both essays presented as handwritten format and as computer-text format;

H_1 : There is difference on raters' score for both essays presented as handwritten format and as computer-text format;

H_0 : The length of essay does not eliminate the presentation effect;

H_1 : The length of essay eliminates the presentation effect.