# A study on ESL writing assessment: Intra-rater reliability of ESL compositions

**Dongwan Cho**
**Pohang University of Science and Technology**

## Abstract

Much research on the inter-rater reliability on ESL writing has been so far conducted to examine consistency across raters and results have shown a great deal of variability. Surprisingly, however, a study focusing on the intra-rater reliability or internal consistency of a single rater is rarely found in the field of ESL writing assessment. The intra-rater reliability of an individual is as significant as inter-rater reliability, since if the former is not secure, neither is the latter. Keeping this concern in mind, this research tries to demonstrate how consistently raters assess ESL compositions over time. For this purpose ten raters who participated in the research as subjects were given twenty essays and rated them in four sessions. The interval between the sessions was about one month or one and half months. In session 1, the raters were asked to rate the essays on their own rating criteria, while in session 2 and session 3, they were instructed to judge the essays based on criteria given to them. In session 4, they again rated the essays in terms of their own assessment criteria. Telephone interviews revealed that the rating criteria offered to them in sessions 2 and 3 were almost the same as their own rating criteria, which made it possible to compare the rating results of each session. The statistical analysis made on descriptive statistics, correlation coefficients and paired *t*-tests showed that the raters of this research were fairly consistent in their ratings over time. Several suggestions for further research are made to help improve understanding of the intra-rater reliability of ESL compositions.

## 1. Introduction

In evaluating ESL compositions, raters' evaluation criteria vary to a great extent across raters. Of the major sub-categories consisting of writing assessment criteria, content and organization have been considered fundamental and important factors. Jacobs and his colleagues' second language writing assessment profile (Jacobs, Zinkgraf, Wormuth, Hartfiel & Hughey, 1981), which has long served

*Melbourne Papers in Language Testing* 1999 Volume 8.1 pp. 1-24. The Language Testing Research Centre, The University of Melbourne.

as the representative assessment criteria for ESL writing, puts the highest weight on content. This profile has been widely used in real rating situations and in teaching ESL composition and evaluation alike. In spite of the popular use of the profile, most raters or teachers seem not to stick to the weight of each category. Instead, they usually apply their own or modified writing assessment criteria to students' compositions. Some place much weight on content, while others do not take into account content or put little emphasis on it. In fact, unlike Jacobs and his colleagues' profile, Bridgeman and Carlson (1983) reported that English language teachers and freshman writing teachers place the quality of content ninth, while organization-related categories such as the whole organization, development of ideas, and paragraph organization are ranked first, second and third, respectively. Another study conducted by Vaughan (1991) supports the above-mentioned point in that raters' comments on writing samples differed markedly, which implies that raters would apply different evaluation criteria to the same writing samples. Application of different assessment criteria is affected by the level of language proficiency of students, the raters' perception of good writing, the course objectives and/or the purpose of the test given to students.

Of several factors leading raters to adopt different evaluation criteria, the rater factor seems to be the most plausible one to account for differences in ratings. Even in cases when the same criteria are provided to raters, there is a great deal of variability among them, because they have different expectations of the same composition derived from their own backgrounds and responses to students' linguistic background (Hamp-Lyons, 1989), and differences in their discipline, sex and amount of exposure to ESL writing (Vann, Lorenz and Meyer, 1991) or just because they were not trained how to apply established criteria to writing samples. Other studies (Lee, 1998; Weir, 1993) focusing on the reliability of writing assessment have also shown that the evaluation made by raters differed greatly. Lee's study conducted in Korea reported a very low inter-rater reliability of .47, .28, and .36 of three groups divided by raters' teaching and grading experience. Weir's study consisting of twenty-two MA students in TESOL also presented an unacceptable inter-rater reliability. A script was rated 5 by one rater and 20 by another one, making a difference of fifteen-points on a twenty-point rating scale.

In the same vein, in a study for assessing rater variability of speaking tests, Mullen (1980) reported differences of at least one point on a five-point rating scale in every pair of judges who were assigned to rate non-native speakers' speaking proficiency. Lumley and McNamara (1995), who adopted a new analytical tool called FACETS, also found a great deal of discrepancy among raters, reporting high reliability of rater separation: .89 in the first round of rating and .87 in the second round of rating. Unlike the traditional concept of inter-rater reliability, the reliability of rater separation signifies consistent patterns of rater disagreement. In other words, the higher the value, the lower the inter-rater reliability.

As shown in the above-mentioned studies, research focusing on rater reliability of the assessment of speaking and writing proficiency generally concluded that there was a great deal of variability across raters. Rating discrepancy between raters may cause a very serious impediment to assuring test validation, thereby incurring the mistrust of the language assessment process itself. The mistrust, or unreliability of testing, in turn, threatens criterion-related evidence of validity of testing, especially a predictive value of a test administered, since assessment results cannot be used for predicting the performance of students in the future.

To reduce unreliability mainly caused by the rater factor, experts have recommended training raters. In an experimental study on effects of training on raters of ESL compositions, Weigle (1994) reported that after a training process, or norming process, several new raters, who were first involved in a writing assessment task, could be in line with the rest of the raters. Other studies also showed positive effects of rater training on writing assessment by helping raters to obtain a clearer concept of the intended rating criteria (Charney, 1984) and to modify expectations of good writing (Huot, 1990). A writing assessment study conducted by Carlson et. al. (1985) reported consistently high inter-rater reliability of Spearman-Brown corrected *r* over .80 after a training session was given to raters. Taken all together, it can be concluded that rater training could make a major contribution to boosting inter-rater reliability, thus leading to acceptable reliability.

Rater-training should be carried out when there is a need for consistency and agreement among raters, as in placement tests, or when the same language classes are provided to students and their language proficiency is evaluated by several teachers in charge of teaching the same classes. In contrast, there are circumstances where language assessment is usually performed by an individual rater or teacher who is solely responsible for teaching and grading. In fact, in many of the classes given at the college level in ESL or EFL settings, a single teacher is usually in charge of evaluating the performance or the achievement of students. This is also true even when the same language class is given by different teachers. In such situations, one of the points to consider is to know how consistently an individual teacher or rater performs his/her rating task. The internal consistency of an individual rater, or intra-rater reliability is as significant as inter-rater reliability, since if the former is not secure, neither is the latter. Surprisingly, however, research on the consistency or reliability of an individual rater or teacher is rarely found in the field of ESL or EFL writing assessment. Given this fact, it is worthwhile to investigate the internal consistency of individual raters. Keeping this concern in mind, this study aims to show whether individual raters are consistent in their ratings over time.

## 2. Method

### 2.1 Subjects

The subjects for this research consisted of 10 raters, all of whom were teaching English at the college or university level in Korea. Out of them, two were native speakers of English, and eight of them were Koreans. All of them had ESL composition teaching experience in Korea or the US, which varied from one year to 10 years. Differences in the subjects' race were not considered a factor to be investigated.

### 2.2 Materials

The materials of this study were short essays written by students who were taking an expository writing class offered at the Pohang University of Science and Technology, Pohang, Korea. At the beginning of the semester, the students taking the class were required to write a short essay of two or three paragraphs. The essays had the same topic, which was 'The development of science and its impact on human life,' and they were typed. Thus such factors as 'title' and 'hand-writing' which might affect writing assessment were eliminated. The number of the essays collected over four semesters was about one hundred. Twenty essays out of the hundred were selected in terms of the researcher's judgment of the level of the essays: very good, good, or poor. The selection on the basis of the writing level of the papers was intentionally made because if all papers with a similar level happened to be selected, it would not represent real situations of a class given to the students. The twenty essays could not represent all the essays in every sense but they were believed to be representative of all the essays in terms of the aforementioned writing levels. Appendix A shows three essays which were rated 'very good,' 'good,' and 'poor' by many of the raters.

### 2.3 Procedures

Data collection of this study took place from September 1998 to February 1999, that is, for about six months. All raters took part in four data collection sessions as shown below:

1) In the first session which took place in September 1998, the raters rated the essays on the basis of their own evaluation criteria.

2) About 1 month later, the same essays which were scanned and formatted in the same fonts were evaluated. The raters were given simple rating guidelines based on a holistic approach (See Appendix B).

3) About 1 month later, the same essays given in session 2 were rated on the basis of discrete-point rating guidelines (See Appendix B).

4) About 1.5 months later, the twenty original essays as given in the first session without any modification were rated using the raters' own criteria.

More detailed data collection procedures are provided below.

In session 1, twenty original essays were rated by one of nine bands: 'very good+,' 'very good,' 'very good-,' 'good+,' 'good,' 'good-,' 'poor+,' 'poor,' 'poor-.' In this session, the raters were instructed to rate the essays on the basis of their own criteria. The reason for adopting these rating bands was that it would reflect the real situations in which writing assessment usually takes place. In other words, in general classroom settings the students' papers are rated in terms of the three bands of 'very good,' 'good,' and 'poor.' The other two extreme bands such as 'excellent' and 'very poor' were not considered here, since in many classes with students of a similar level of language proficiency 'excellent' and 'very poor' are rarely given to students. The raters were also asked to read the essays and finish rating them on the same day as they started. This was intended to exclude a possibility that any inconsistency in rating may occur when it takes more than one day. After the rating task was over, the raters mailed the rating results along with the essays. Thus they were not allowed to keep the essays with them.

In session 2, the same twenty essays which had been scanned and formatted in the same fonts were mailed to the raters one month later. Having them scanned and formatted in the same fonts was intended not to give any impression to the raters that the same essays would be rated, which may affect ratings of this session by reminding themselves of the rating results of session 1. Additionally, the arranged order of the essays was changed for the same purpose. In this session, the raters were instructed to rate the essays on the basis of the rating guidelines which had much resemblance to the holistic assessment guidelines developed by Jacobs' and his colleagues (Appendix B). When adopting the rating guidelines, the researcher assumed that they were popularly used and thus were quite similar to the guidelines to which the raters applied in session 1. If the rating guidelines the raters had applied in session 1 were different from those provided in session 2, then investigating the consistency or reliability of individual raters did not make any sense. In fact, a telephone interview was carried out 1) to check whether the guidelines provided to the raters were the same as or similar to raters' own rating criteria, 2) whether they could recognize that the essays given to them this time were exactly the same as those rated in session 1, and 3) to demonstrate whether they could remember and tried to duplicate the results of their ratings in session 1 and session 2.

About a month later, the raters were given the same twenty essays with discrete-point rating guidelines. The purpose of this session was to compare the rating results based on the discrete-point approach and those based on the holistic approach used in session 2. For this purpose, another trick was made. To distract the raters' attention, the raters were asked to rate several major categories such as structure, language use, vocabulary and mechanics along with giving an overall rating to each essay. If the same holistic grading guidelines provided to them for session 2 were given to them in this session, they would think that they were doing the same task. Thus a way to lead them to think that they were doing a different task was needed, which was to provide them with discrete-point rating guidelines. Comparisons were made between the rating results of session 2 and the overall scores of session 3. After the results of this session have been collected, a telephone interview was given to illustrate 1) whether there were any differences between the rating scales of session 2 and those of session 3, 2) how raters came up with the overall scores, and 3) whether the results of session 2 affected those of this session.

About one and half months later, the data for the last session were collected. As in session 2, the purpose of this session was again to confirm the rating consistency or reliability of individual raters. For this purpose, the same twenty original essays which had been given to the raters as in session 1 were rated on the basis of the raters' own judgement. A telephone interview was given to show 1) whether they could retrieve the results of their ratings of the previous sessions and 2) whether they tried to replicate the rating results.

## 3. Results

Several statistical analyses were made to demonstrate the intra-rater reliability of ratings over time.

### 3.1 Intra-rater reliability between session 1 and session 2

Table 1 below shows descriptive statistics of differences of ratings based on a 9-band scale between session 1 and session 2. In order to conduct quantitative analysis 'very good,' 'good,' and 'poor' were converted into numeric values: 'very good+' into '9,' 'very good' into '8,' 'very good-' into '7,' 'good+' into '6,' 'good' into '5,' 'good-' into '4,' 'poor+' into '3,' 'poor' into '2,' and 'poor-' into '1.'

| Rater | Number of essays | Absolute value of point difference between session 1 and session 2 | | | | | | Proportion of 0 and 1 point difference |
|---|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 | 5 | |
| 1 | 20 | 4 | 9 | 6 | 1 | 0 | 0 | 0.65 |
| 2 | 20 | 4 | 7 | 5 | 3 | 1 | 0 | 0.55 |
| 3 | 20 | 10 | 9 | 1 | 0 | 0 | 0 | 0.95 |
| 4 | 20 | 7 | 7 | 4 | 1 | 1 | 0 | 0.70 |
| 5 | 20 | 5 | 8 | 4 | 2 | 1 | 0 | 0.65 |
| 6 | 20 | 6 | 11 | 2 | 0 | 1 | 0 | 0.85 |
| 7 | 20 | 6 | 6 | 2 | 3 | 2 | 1 | 0.60 |
| 8 | 20 | 6 | 6 | 3 | 1 | 3 | 1 | 0.60 |
| 9 | 20 | 8 | 7 | 3 | 2 | 0 | 0 | 0.75 |
| 10 | 20 | 8 | 9 | 2 | 1 | 0 | 0 | 0.85 |

**Table 1: Descriptive statistics of differences in ratings between session 1 and session 2**

Since there is no established criteria for determining acceptable intra-rater reliability, it was necessary to establish the extent to which the differences of ratings made between session 1 and session 2 within a month interval could be considered consistent and reliable. Considering the possibility of discrepancy of ratings of an individual rater or teacher which may occur in real writing assessment situations, it was assumed that one point difference was reliable and acceptable. Converting the sum of no difference and 1 point difference between session 1 and session 2 into a proportion of the number of all essays shows .65 for rater 1, .55 for rater 2, .95 for rater 3, .70 for rater 4, .65 for rater 5, .85 for rater 6, .60 for rater 7, .60 for rater 8, .75 for rater 9 and .85 for rater 10, respectively. If we set up .7 as a cut-off value for reliable and consistent intra-rater reliability, we can say that five out of ten raters, 3, 4, 6, 9 and 10 showing over .7 of internal consistency turned out to be consistent in their ratings.

Another statistical method adopted to demonstrate intra-rater reliability is the coefficient alpha, or Cronbach's alpha and Kendall's tau-b correlation coefficients. As Bachman (1991) recommended, the coefficient alpha was obtained to illustrate internal consistency of two ratings over time for each rater. In addition, Kendall's tau-b correlation coefficients, which demonstrate the correlation based on the number of concordant and discordant pairs of observations (SAS Users' Guide: Version 6), were calculated. Kendall's tau-b coefficient will show a smaller value of correlation coefficients, compared to Cronbach's alpha. Table 2 below shows Cronbach's alpha and Kendall's tau-b coefficients between session 1 and session 2.

| Rater | Cronbach's coefficient alpha | Kendall's tau-b coefficient |
|---|---|---|
| 1 | 0.92 | 0.75 |
| 2 | 0.81 | 0.61 |
| 3 | 0.96 | 0.84 |
| 4 | 0.90 | 0.70 |
| 5 | 0.90 | 0.70 |
| 6 | 0.86 | 0.67 |
| 7 | 0.92 | 0.78 |
| 8 | 0.52 | 0.27 |
| 9 | 0.91 | 0.65 |
| 10 | 0.97 | 0.88 |

**Table 2: Cronbach's coefficient alpha and Kendall's tau-b coefficients between session 1 and session 2**

As noticed in the above table, Cronbach's coefficient alpha is very high across the raters except for rater 8, which seems to overestimate the internal consistency of the raters. If we apply .7 as an acceptable value for intra-rater consistency, nine out of ten raters were said to be highly consistent in their ratings, which is not congruent with the findings derived from descriptive statistics. In contrast, Kendall's tau-b coefficients, which report moderate correlation coefficients seem more appropriate in showing the intra-rater reliability of each rater. If we set up .7 as a cut-off point for demonstrating acceptable internal consistency of the raters, raters 1, 3, 4, 5, 7 and 10 reached the value. In other words, six out of ten raters were found to be consistent in their ratings. Even though raters 6 and 9 do not meet .7 of Kendall's tau-b coefficients, theirs are quite close to the set-up value, which show .67 and .65, respectively.

Along with descriptive statistics and two correlation coefficients, the paired *t*-test with data of sessions 1 and 2 was carried out to illustrate the consistency of ratings. When conducting the test, the difference of each pair of the essays was changed into an absolute value, since the purpose of this study was to demonstrate the consistency of ratings over time, not to compare the effect of a factor to ratings of following sessions. In other words, to know whether the rating results of session 1 were bigger or smaller than those of session 2 was not a concern of this study. This study was interested in only the absolute differences of ratings between the sessions. Table 3 shows the statistical results of the paired *t*-test of session 1 and session 2.

| Rater | Mean | Std Dev | Minimum | Maximum | *T* | Prob>\|*T*\| |
|---|---|---|---|---|---|---|
| 1 | 1.20 | 0.83 | 0 | 3.00 | 6.44 | 0.0001 |
| 2 | 1.50 | 1.15 | 0 | 4.00 | 5.85 | 0.0001 |
| 3 | 0.55 | 0.60 | 0 | 2.00 | 4.07 | 0.0007 |
| 4 | 1.10 | 1.12 | 0 | 4.00 | 4.40 | 0.0003 |
| 5 | 1.30 | 1.13 | 0 | 4.00 | 5.15 | 0.0001 |
| 6 | 0.95 | 0.94 | 0 | 4.00 | 4.50 | 0.0002 |
| 7 | 1.60 | 1.57 | 0 | 5.00 | 4.56 | 0.0002 |
| 8 | 1.60 | 1.60 | 0 | 5.00 | 4.46 | 0.0003 |
| 9 | 0.95 | 1.00 | 0 | 3.00 | 4.25 | 0.0004 |
| 10 | 0.80 | 0.83 | 0 | 3.00 | 4.30 | 0.0004 |

**Table 3: Results of the paired *t*-test of session 1 and session 2**

If we consider the mean of the differences less than 1 to be consistent, we can state that raters 3, 6, 9 and 10 were consistent in their assessment over time. And since the probability of obtaining the mean of the differences for these raters by chance is .0007 for rater 3, .0002 for rater 6, .0004 for rater 9 and .0004 for rater 10, respectively, we can state that the mean is significantly meaningful.

### 3.2 Analysis of telephone interview given after session 2

A telephone interview was given 1) to check whether the guidelines provided to the raters in session 2 were the same as or similar to the raters' own rating criteria they adopted in session 1, 2) to check whether they could recognize that the essays given to them in session 2 were exactly the same as those rated in session 1, and 3) to demonstrate whether they could remember the rating results and tried to duplicate the rating results of session 1 and session 2.

In answer to question 1 above, eight out of the ten raters responded that the guidelines given to them in session 1 and their own rating criteria were almost the same. One rater, rater 4, answered that the guidelines provided in session 2 were more strict than her own criteria, since they contained such sub-categories as grammar and mechanics which she had not considered in session 1. In contrast, rater 6 responded that her own criteria were more strict than the guidelines provided to her in session 2 and thus she seemed to have given poorer ratings especially to low level essays. Both of them, however, answered that there was not much difference between the two rating criteria. With respect to the next question concerning whether the raters realized that the essays given to them in session 1 and session 2 were the same, seven raters answered that they did not realize that the essays were the same. Here it should be mentioned that all of the raters thought that the essays given to them in session 2 were modified ones and thus looked better than those in session 1. Two raters, rater 2 and rater 7, stated that the essays looked the same even though they were not sure that the essays were exactly the same. Only one rater, rater 3 answered that she found that they were exactly the same. The next question was concerned with whether they could remember their rating results of session 1 and tried to duplicate the results of the two sessions. In answer to this question, all of them answered that they could not remember exactly the rating results of session 1 and never tried to replicate their rating results of the two sessions. However, some of them pointed out that they might have a vague idea of the rating results of the essays which fell into the two extreme bands such as 'very good+' and 'poor-' Even in the case, however, they said they could not recall exact rating results of session 1. To sum up the telephone interview, it can be said that the raters adopted very similar rating criteria both in session 1 and session 2 and the rating results of session 1 did not affect the ratings of session 2. This provides a solid ground for making it possible to compare the rating results of session 1 and session 2.

### 3.3 Intra-rater reliability between session 2 and session 3

Several statistical analyses to illustrate internal consistency of ratings between session 2 and session 3 were made. In session 3, the raters were given discrete-point rating guidelines with four sub-categories such as structure, language use or grammar, vocabulary and mechanics. Added to these categories, they were asked to determine the overall rating of each essay. Table 4 shows descriptive statistics of differences in ratings between session 2 and session 3. The ratings of session 3 were based on the overall scores.

| Rater | Number of essays | Absolute value of point difference between session 2 and session 3 | | | | | | Proportion of 0 and 1 point difference |
|---|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 | 5 | |
| 1 | 20 | 4 | 12 | 1 | 3 | 0 | 0 | 0.80 |
| 2 | 20 | 3 | 7 | 4 | 5 | 0 | 1 | 0.50 |
| 3 | 20 | 11 | 8 | 1 | 0 | 0 | 0 | 0.95 |
| 4 | 20 | 10 | 6 | 3 | 1 | 0 | 0 | 0.80 |
| 5 | 20 | 7 | 8 | 5 | 0 | 0 | 0 | 0.75 |
| 6 | 20 | 9 | 9 | 2 | 0 | 0 | 0 | 0.90 |
| 7 | 20 | 13 | 5 | 2 | 0 | 0 | 0 | 0.90 |
| 8 | 20 | 11 | 7 | 2 | 0 | 0 | 0 | 0.90 |
| 9 | 20 | 11 | 7 | 1 | 1 | 0 | 0 | 0.90 |
| 10 | 20 | 13 | 7 | 0 | 0 | 0 | 0 | 1.00 |

**Table 4: Descriptive statistics of differences in ratings between session 2 and session 3**

As seen in table 4, the number of the essays which illustrates no difference and 1-point differences between session 2 and session 3 is 16 for rater 1, 10 for rater 2, 19 for rater 3, 16 for rater 4, 15 for rater 5, 18 for rater 6, 18 for rater 7, 18 for rater 8, 18 for rater 9 and 20 for rater 10. The right column of table 4 shows the proportion of the number of the essays with 0 and 1-point differences out of the twenty essays. Except for rater 2, all nine raters showed a high level of internal consistency or intra-rater reliability. Compared to the intra-rater reliability between session 1 and session 2, that of session 2 and session 3 is much higher. This high level of internal consistency based on descriptive statistics is supported by Kendall's tau-b coefficients which report a consistently high level of correlation coefficients.

| Rater | Cronbach's coefficient alpha | Kendall's tau-b coefficient |
|---|---|---|
| 1 | 0.93 | 0.73 |
| 2 | 0.79 | 0.46 |
| 3 | 0.96 | 0.82 |
| 4 | 0.88 | 0.71 |
| 5 | 0.91 | 0.87 |
| 6 | 0.92 | 0.70 |
| 7 | 0.89 | 0.90 |
| 8 | 0.92 | 0.74 |
| 9 | 0.87 | 0.84 |
| 10 | 0.92 | 0.93 |

**Table 5: Cronbach's coefficient alpha and Kendall's tau-b coefficients between session 2 and session 3**

Kendall's tau-b coefficients show that only one rater, rater 2, failed to show acceptable internal consistency of ratings. Five out of the ten raters illustrated over .8 of the coefficients.

| Rater | Mean | Std Dev | Minimum | Maximum | *T* | Prob>\|*T*\| |
|---|---|---|---|---|---|---|
| 1 | 1.10 | 0.85 | 0 | 3.00 | 5.77 | 0.0001 |
| 2 | 1.75 | 1.29 | 0 | 5.00 | 6.05 | 0.0001 |
| 3 | 0.50 | 0.61 | 0 | 2.00 | 3.68 | 0.0016 |
| 4 | 0.85 | 0.93 | 0 | 3.00 | 4.07 | 0.0006 |
| 5 | 0.90 | 0.79 | 0 | 2.00 | 5.11 | 0.0001 |
| 6 | 0.65 | 0.67 | 0 | 2.00 | 4.33 | 0.0004 |
| 7 | 0.45 | 0.69 | 0 | 2.00 | 2.93 | 0.0086 |
| 8 | 0.55 | 0.69 | 0 | 2.00 | 3.58 | 0.0020 |
| 9 | 0.60 | 0.82 | 0 | 3.00 | 3.27 | 0.0040 |
| 10 | 0.35 | 0.49 | 0 | 1.00 | 3.20 | 0.0047 |

**Table 6: Results of the paired *t*-test of session 2 and session 3**

The results of the paired *t*-test also support the above-mentioned point in that only two raters, rater 1 and rater 2, showed greater than 1 of the mean of the differences for each essay.

### 3.4 Analysis of telephone interview given after session 3

Another telephone interview was given after session 3. The interview consisted of several questions in order to illustrate 1) whether there were any differences between the rating scales of session 2 and those of session 3, 2) how raters came up with the overall scores, and 3) whether the results of session 2 affected those of session 3.

The responses to question 1 above varied among the raters. Seven raters answered that there was not much difference in rating guidelines between session 2 and session 3 even though the guidelines of session 3 were more detailed and explicit. Two raters mentioned that since more detailed guidelines for each category were given to them in this session, rating was more difficult than session 2 in which they were instructed to rate the essays on the basis of holistic rating guidelines. In contrast, only one rater, rater 4, responded that ratings of this session were easier than session 2 because of the same reason. In answer to question 2, how they came up with the overall score of each essay, nine raters answered that they gave their ratings based on their holistic impression of each essay along with considering some sub-categories. The fact that most of the raters applied similar rating guidelines in session 2 and session 3 made it possible to compare the ratings of session 2 and session 3. With respect to question 3, as to whether the rating results of session 2 affected those of session 3, all raters answered that the former did not affect the latter even though some of them could remember the ratings of the essays in the extreme bands such as 'very good+' and 'poor-.'

### 3.5 Intra-rater reliability between session 1 and session 4

In session 4, the raters were instructed to rate the same twenty original essays on their own rating criteria. Thus conditions for ratings between session 1 and session 4 were believed to be the same except for a possibility that the raters could remember the rating results of the previous sessions. There was about a six-month interval between session 1 and session 4.

| Rater | Number of essays | Absolute value of point difference between session 1 and session 4 | | | | | | Proportion of 0 and 1 point difference |
|---|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 | 5 | |
| 1 | 20 | 8 | 8 | 4 | 0 | 0 | 0 | 0.80 |
| 2 | 20 | 6 | 6 | 6 | 1 | 1 | 0 | 0.60 |
| 3 | 20 | 7 | 9 | 3 | 1 | 0 | 0 | 0.80 |
| 4 | 20 | 2 | 10 | 6 | 2 | 0 | 0 | 0.60 |
| 5 | 20 | 7 | 8 | 3 | 2 | 0 | 0 | 0.75 |
| 6 | 20 | 6 | 11 | 3 | 0 | 0 | 0 | 0.85 |
| 7 | 20 | 12 | 6 | 2 | 0 | 0 | 0 | 0.90 |
| 8 | 20 | 12 | 5 | 0 | 2 | 1 | 0 | 0.85 |
| 9 | 20 | 5 | 10 | 5 | 0 | 0 | 0 | 0.75 |
| 10 | 20 | 9 | 8 | 2 | 1 | 0 | 0 | 0.85 |

**Table 7: Descriptive statistics of differences in ratings between session 1 and session 4**

Following the guidelines for acceptable intra-rater reliability set up in this study, we can state that eight raters out of ten proved to be consistent in their ratings over time. Only two raters, rater 2 and rater 4, illustrated slightly low internal consistency in their ratings, .60 for both of them. This finding is clearly supported by Kendall's tau-b correlation coefficients.

| Rater | Cronbach's coefficient alpha | Kendall's tau-b coefficient |
|---|---|---|
| 1 | 0.95 | 0.77 |
| 2 | 0.82 | 0.58 |
| 3 | 0.93 | 0.74 |
| 4 | 0.87 | 0.67 |
| 5 | 0.94 | 0.75 |
| 6 | 0.93 | 0.77 |
| 7 | 0.98 | 0.90 |
| 8 | 0.87 | 0.74 |
| 9 | 0.90 | 0.67 |
| 10 | 0.96 | 0.89 |

**Table 8: Cronbach's coefficient alpha and Kendall's tau-b coefficients between session 1 and session 4**

Similar to the findings illustrated in the descriptive statistics, three raters, raters 2, 4 and 9, did not show acceptable internal consistency of ratings. The correlation coefficients of the three raters, however, are not low enough to claim that they rated the essays inconsistently. Kendall's tau-b coefficient of rater 4 and rater 9 is very close to the acceptable value of intra-rater reliability, showing .67.

| Rater | Mean | Std Dev | Minimum | Maximum | *T* | Prob>\|*T*\| |
|---|---|---|---|---|---|---|
| 1 | 0.80 | 0.77 | 0 | 2.00 | 4.66 | 0.0001 |
| 2 | 1.25 | 1.12 | 0 | 4.00 | 5.00 | 0.0001 |
| 3 | 0.90 | 0.85 | 0 | 3.00 | 4.72 | 0.0001 |
| 4 | 1.40 | 0.82 | 0 | 3.00 | 7.63 | 0.0001 |
| 5 | 1.00 | 0.79 | 0 | 3.00 | 4.59 | 0.0002 |
| 6 | 0.85 | 0.67 | 0 | 2.00 | 5.67 | 0.0001 |
| 7 | 0.50 | 0.69 | 0 | 2.00 | 3.25 | 0.0042 |
| 8 | 0.75 | 1.21 | 0 | 4.00 | 2.78 | 0.0121 |
| 9 | 1.00 | 0.73 | 0 | 2.00 | 6.16 | 0.0001 |
| 10 | 0.75 | 0.85 | 0 | 1.00 | 3.94 | 0.0009 |

**Table 9: Results of paired *t*-test of session 1 and session 4**

As seen in table 9, raters 2 and 4 showed greater than 1 of the mean of the differences between session 1 and session 4. In contrast, the mean of the differences of the other eight raters is less than 1, which implies that the raters were consistent and reliable in their ratings. In addition, since the probability of the mean of the differences of the eight raters, or the *t* value, is very low, we can say that the mean differences are significantly meaningful.

### 3.6 Analysis of telephone interview given after session 4

A telephone interview was given to show 1) whether raters could retrieve the results of their ratings of the previous sessions and 2) whether they tried to replicate the rating results. In answer to question 1, seven raters answered that they could not remember the results of their ratings of the previous sessions. In contrast, three pointed out that they could recall the general bands of their ratings of the previous sessions, for example, 'very good,' 'good' and 'poor.' All mentioned that they never tried to duplicate their ratings results.

## 4. Discussion

The statistical analyses made above show somewhat unexpected results in that most of the raters kept internal consistency in their ratings. Since there has been little research on intra-rater reliability of ESL writing, it is not known exactly how consistently raters evaluate ESL compositions over time. Teachers and experts in ESL testing as well seem to cast doubt on the internal consistency of raters. It is jokingly mentioned that the rating in the morning may be different from that in the evening on the same day. Unlike this unsupported conjecture, most of the raters involved in this study turned out to be consistent and reliable in their ratings. Out of the three comparisons made between session 1 and session 2, session 2 and session 3, and session 4 and session 1, ratings between session 2 and session 3 were most consistent among the comparisons between the sessions. This is because the raters adopted very similar rating criteria in session 2 and session 3, respectively. In session 2, they were instructed to rate the essays in terms of a holistic rating scale, while in session 3 they were given similar assessment criteria. According to table 4, nine of ten raters illustrated over .7 of the proportion of 0 and 1-point difference. This high rating consistency is supported by a high level of correlation coefficients of Kendall's tau-b, which reported that only one rater out of ten did not come up with acceptable internal consistency. In contrast, the comparisons made between session 1 and session 2 demonstrated the least consistent phase in ratings. Only five raters showed over .7 of the proportion of 0 and 1-point difference among the twenty essays compared. In session 1 the raters were instructed to rate the essays on their own rating criteria, whereas in session 2 they rated the essays on the basis of a holistic rating criteria provided to them. This low intra-rater reliability between session 1 and session 2 may be due to the fact that raters would apply different rating criteria, even though in the phone interview most of them replied that there was not much difference between their own rating criteria in session 1 and ones which were provided to them for session 2. This low intra-rater reliability, however, cannot lead us to conclude that the raters were inconsistent in their ratings, since comparisons made between session 4 and session 1 showed a high level of internal consistency. In the comparisons eight raters demonstrated over .7 of the proportion of 0 and 1-point difference. Added to that, Kendall's tau-b coefficients reported that only three raters failed to show acceptable internal reliability. Even in the case, however, two raters' coefficients were .68, which was very close to .70. The findings based on descriptive statistics and correlation coefficients can lead us to claim that most of the raters involved in the study were highly consistent in their ratings of ESL compositions.

## 5. Suggestions for further research

Since this study targeted the intra-rater reliability of ESL compositions, the rater was placed at the center of this research. The judgment of internal consistency was based on, for example, how many raters demonstrated correlation coefficients of over .7. The interpretation on the group level, however, did not show the general characteristics of an individual rater's rating behaviors. One factor which can be investigated here is the amount of ESL writing teaching experience and its impact on rating behaviors. The raters who participated in this study had at least one year experience of teaching ESL composition. It was found that the amount of teaching experience of raters did not have any direct relation to the consistency of ratings. Rater 2, who turned out to be least consistent in ratings, had comparatively longer teaching experience. In contrast, rater 3, who showed a high level of intra-rater reliability, had one year of ESL composition teaching experience. Thus we would say that teaching experience seems to be irrelevant to the internal consistency of ratings. Rather, it can be conjectured that such factors as raters' personality and/or different backgrounds seem to be more relevant to ratings. If research focusing on these factors is conducted, several new findings of inter-rater reliability may be revealed.

Along with the rater factor, the level of the essays can be pointed out as a factor which may influence the results of this study. As mentioned before, in this study the researcher chose twenty essays out of one hundred essays written by students who took an ESL writing class the researcher offered. Selecting the essays was made intentionally in terms of the researcher's judgement of writing level since the researcher was worried that if the essays displaying a similar level happened to be selected, it might not represent real classroom situations. The twenty essays could not exactly represent all the essays but they were believed to be representative of all the essays with respect to their writing levels. The findings of this research, however, may not be supported if essays showing a similar writing proficiency were provided to the raters. If a researcher conducts the same research as this study with essays displaying a similar writing proficiency, raters may evince different behaviors of ratings. In other words, if essays showing a similar level of writing proficiency are rated in terms of a 9-point rating scale as the raters of this study actually did, raters are likely to demonstrate lower levels of intra-rater reliability, because the levels of the essays would not be easily distinguishable.

Lastly, another factor which may influence the internal consistency of ratings is a memory factor. Even though in the telephone interviews most of the raters answered that they never tried to duplicate the rating results, it is quite probable that the raters' memory of the previous sessions had an effect on the ratings of the sessions that followed. And in fact, some of the raters, especially in the interview given after session 4, mentioned that they could recall some of their ratings of the previous sessions. These raters were found to be fairly reliable and consistent in their ratings. These finding make us claim that the raters' memory influenced subsequent ratings. One way to get rid of this possibility is to lengthen the interval between the sessions of the research. Even in this situation, however, we are not sure whether memory would not have any impact on the ratings of following sessions.

If we take into account the factors mentioned above and invent ways to control them, we can have a better understanding of the intra-rater reliability of ESL compositions. As mentioned in the introductory statement, it should be again addressed that the internal consistency of the ratings of ESL compositions is as important as inter-rater reliability, since if the former is not secure, neither is the latter.

Acknowledgements: The author wishes to acknowledge the financial support of the Korea Research Foundation in the program of 1997.

## 6. References

Bachman, L. (1991). *Fundamental considerations in language testing.* Hong Kong: Oxford University Press.

Bridgeman, B., & Carlson, S. (1983). *Survey of academic writing tasks required of graduate and undergraduate foreign students.* Princeton, NJ: Educational Testing Service.

Carlson, S., Bridgeman, B., Camp, R., & Waanders, J. (1985). *Relationship of admission test scores to writing performance of native* *and nonnative speakers of English* (TOEFL Research Report No. 19). Princeton, NJ.: Educational Testing Service.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the Teaching of English, 18* (1), 65-81.

Hamp-Lyons, L. (1989). Raters respond to rhetoric in writing. In H. Dechert and G. Raupach (Eds.), *Interlingual processes* (pp. 229-244). Tubingen: Gunter Narr Verlag.

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research, 60*, 237-263.

Jacobs, H., Zinkgraf, S., Wormuth, D., Hartfiel, V., & Hughey, J. (1981). *Testing ESL composition: A practical approach.* Rowley, MA: Newbury House.

Lee, Y. (1998). An investigation into Korean markers' reliability for English writing assessment. *English Teaching, 53* (1), 179-200.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12* (1), 54-71.

Mullen, K. (1980). Evaluating writing proficiency in ESL. In J. Oller and K. Perkins (Eds.), *Research in language testing* (pp. 91-101). Rowley, Mass: Newbury House.

SAS Institute, Inc. (1990). *SAS user's guide: Statistics.* Cary, NC: Author.

Vann, R. J., Lorenz, F., & Meyer, D. (1993). Error gravity: Faculty response to errors in the written discourse of nonnative speakers of English. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp.181-195). Norwood, NJ: Ablex Publishing Corporation.

Vaughan, C. (1993). Holistic assessment: What goes on in the rater's mind? In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp.111-125). Norwood, NJ: Ablex Publishing Corporation.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11* (2): 197-223.

Weir, C. (1993). *Understanding and developing language tests.* New York: Prentice Hall.

## Appendix A: Three representative writing samples

### The development of science and its impact on human society

It wasn't until a few decades ago that the science really started to change the contour of society. But once it got going, it didn't just go forward; it sprinted forward, Since the scientific revolution probably the most dramatic changes in human society in about 3000 years of human hi have occurred. The application of scientific principles has produced the automobiles, TV, fluorescent lights, etc. These items have now become a part of our lives; one would find it extremely inconvenient to live without any one of these even for a day. Not that every 'invention was great. The development of nuclear weapons certainly makes us wonder if sticking with science is really good for our health. This is why the science should be guided. Because, while the science can certainly make our lives more comfortable and safe, any uncontrolled growth of science can be dangerous. President Clinton's policy of banning the cloning of human beings is a good example. In other words, we should sometimes intervene and put the science on the correct path.

The science has not only made our lives more comfortable but it has considerably altered the way we view the world. We no longer say things like 'The stars glitter in the night sky because God made it glitter, the ivy twines because God make it twine and the sky is blue because God made it blue.' Today we view the world mechanically. In such a view the glittering stars, twining ivies and the blue sky are nothing more than natural consequences of simple cause-and-effect. Some people get depressed when they learn these truths. They are dismayed upon the fact that the planet we live in is only a tiny speck of dust rotating far out on one spoke of Milky Way galaxy. I don't blame them. If they feel so, its because we have been too busy doing science to care about such eternal verities as honor, love and patriotism. Therefore we should try to preserve our tradition and culture as we do the science.

The potential of the science is undoubtedly limitless. But we must always keep in mind that it's our responsibility to leave our future generations with clean and safe environment. If we are not careful every time we take a step forward, we might well wake up one morning to discover the earth turned into one big dead planet. (Rated 'Very Good' by most of the raters)

**Does science improve human life?**

Development of science changes human life in various sides. It brings us to go abroad by airplane and shop in suburb by car. We have many kinds of goods that would have been not, such as watch, cassette tape player. We also reduce much of amount of chores by using machines or computers. Biology and medical technology make it possible to cure of such disease which once impossible. Scientific developments make us do once impossible, have something was not, reduce our work, and save our health or life.

It may be obvious science improve human life. But human life is not a something that is evaluated only by doing, having. Not having little, one can be happy in other words, he/she can have good life. It can not be said that long life is happier than short life.

Moreover some scientific developments do negative in human life. Science makes it possible more powerful weapon kill more people. It pollutes our environment. Most serious result of the pollution, some plant or animal are disappearing in this planet. We are threaten our life. As a result of its development, social structure changes broke warm relationship among neighbors.

But these are two sides of science. It can not have intention. Human can only have intention. It is up to us whether science beneficial our life or not. As a scientist and an engineer, we must not forget science should be of human. It should work for human life. We are one of human. (Rated 'Good' by most of the raters)

**The development of science and it's effect on human's life**

Science has been developing for human's life. The result is fantastic. Many disease are controlled by science. And beast is not fearful to human. We can live in warm house. We are not hungry anymore.

But I doubt science's worth now. Some days ago the North Korea do missile experiment. Is that science for human? Why science have to develop? To kill people and live fun life for 3 or 4 man? It is the big gamble and dangerous gamble also.

If science is dangerous to people worth to exist is nothing. That science is more good not to exist in the word. We must remember the science have to be the science for human only. This conscience can be born by education. We must not ignore the power of the education. The emotional for human from childhood is only way today's science becomes the science for human.

I believe science can make our life more various. When science is used have good object. (Rated 'Poor' by most of the raters)

## Appendix B

### Directions for rating given in session 2

1) Read the essays as quickly as you can and rate them on the basis of the rating scales given below. There are nine scales for rating: very good+, very good, very good-, good+, good, good-, poor+, poor, poor-. After you have finished rating, please mark the rating results on the rating sheet. In addition, please write down how long it took to read and finish rating each essay.

*Rating guidelines*

Very Good
Writers communicate very effectively. Ideas are expressed clearly and fluently. Vocabulary, sentences and mechanics work effectively to convey the intended ideas.

Good
Writers communicate well. Main ideas are loosely organized but they stand out. Incomplete mastery of some of the criteria for vocabulary, language use, and mechanics limit the writer's effectiveness although the flow of ideas is not seriously impeded.

Poor
Writers communicate partially. On the whole ideas are confused and

disconnected. Lack of mastery of most of the criteria for vocabulary, language use, and mechanics severely restricts the flow of ideas.

2) If possible, finish your rating task on the same day as you start.

3) Some essays do not have any topic. In that case, please regard the topic of the essay as 'The development of science and its impact on human life.'

4) Please return the essays along with the rating sheet.

**Direction for rating given in session 3**

1) Read the essays as quickly as you can and rate them on the basis of the rating scales given below. There are nine scales for rating: very good+, very good, very good-, good+, good, good-, poor+, poor, poor-. After you have finished rating, please mark the rating results on the rating sheet. In addition, please write down how long it took to read and finish rating each essay.

*Rating guidelines*

Structure (Organization)

Very Good: Ideas clearly stated and supported,
well-organized, logical sequence

Good: Ideas are loosely organized but main ideas stand out

Poor: Ideas confused or disconnected, lacks logical sequencing and development

Language Use

Very Good: Few errors of sentence structure, tense, number, articles and prepositions

Good: Some errors of sentence structure, tense, number, articles and prepositions

Poor: Frequent errors of sentence structure, tense, number, articles and prepositions, meaning confused and obscured

Vocabulary

Very Good: appropriate, sophisticated and effective use vocabulary

Good: occasional errors in using vocabulary, but meaning not obscured

Poor: Frequent errors in using vocabulary

Mechanics

Very Good: mastery of conventions, few errors of punctuation, paragraphing, capitalization

Good: occasional errors of punctuation, paragraphing, capitalization

Poor: Frequent errors of punctuation, paragraphing, capitalization

2) If possible, finish your rating task on the same day as you start.

3) Some essays do not have any topic. In that case, please regard the topic of the essay as 'The development of science and its impact on human life.'

4) Please return the essays along with the rating sheet.