Teaching Experience and Evaluation of Second-Language Students' Writing

Ling Shi
University of British Columbia

Wenyu Wang, and Qiufang Wen
Nanjing University, China

This study explores the relationship between teachers' evaluations of second-language (L2) writing and their years of teaching L2 writing. Forty-six English teachers from twenty-three tertiary institutions in Mainland China holistically evaluated ten essays written by Sinophone English majors and justified their scores for each essay with three qualitative comments. Results show that the most experienced writing teachers gave significantly lower scores than did the less or the least experienced writing teachers for four of the ten essays. Analyses of the qualitative comments on these four essays suggest that the experienced writing teachers made either more negative or fewer positive comments on aspects such as *general organization*, *language fluency, ideas* and *general language*.

Cette étude explore les rapports entre l'évaluation de la langue écrite par les enseignants de langue seconde et leur nombre d'années d'expérience en enseignement de la composition. Quarante-six enseignants de vingt-trois établissements d'enseignement supérieur de Chine continentale ont évalué de manière globale dix essais écrits en anglais par des étudiants qui se spécialisent en anglais. Ils ont ensuite justifié leurs notes avec trois commentaires qualitatifs. Les résultats montrent que pour quatre des dix essais, les enseignants les plus expérimentés en enseignement de la composition ont attribué des notes plus basses que leurs collègues moins expérimentés. Une analyse des commentaires qualitatifs sur ces quatre essais indique que les enseignants plus expérimentés ont fait plus de commentaires négatifs ou moins de commentaires positifs sur des aspects tels que *l'organisation générale*, *la fluidité de la langue*, *les idées* et *la qualité générale de la langue*

Introduction

Many English teachers believe that their experience or years of teaching writing and the types of students they teach influence how they evaluate student writing. However, little empirical research has been conducted to lend credence to this teacher belief. In an effort to fill the gap, we conducted the present study in

Address for correspondence: Ling Shi, Faculty of Education, Department of Language and Literacy Education, 2034 Lower Mall Road, University of British Columbia, Vancouver, BC, V6T 1Z2. E-mail: ling.shi@ubc.ca.

Mainland China by asking 46 English teachers with varying years of teaching L2 writing to assess ten essays written by Sinophone English majors. By comparing how participating teachers rated and commented on the essays, the study explored whether the number of years of teaching L2 writing had an impact on participants' evaluations. In this paper, we first review the relevant research that suggests a connection between teachers' experience and their evaluation of L2 writing. We then describe the participating teachers and how their evaluations of students' writing were collected and compared. This is followed by a report of the findings and a discussion which highlights how teachers with varying years of teaching experience in L2 writing were more harsh or lenient in scoring and were more positive or negative for certain language and rhetorical features in students' writing. We conclude by emphasizing research possibilities in the same direction.

Previous studies

To the best of our knowledge, no previous research has examined teachers' experience in teaching L2 writing as a variable that might determine differences in their evaluations of student writing. Research has, however, suggested that raters' general English teaching experience in ESL can have an impact on writing evaluation. In a study that compared raters' criteria for error gravity, Hughes and Lascaratou (1982) noted that, compared with raters with no teaching experience who may depend almost exclusively on the criterion of intelligibility, experienced English teachers tend to make use of the criteria of both intelligibility and grammar rules. In another study, Cumming (1990) compared the decision-making behaviours used by experienced and novice teacher-raters in evaluating ESL writing and found that expert teachers, compared with novices, used more efficient strategies and a wider range of knowledge sources to read and judge students' texts. For instance, the expert teachers attended frequently to certain features such as key criteria, number of main ideas, development of the topic and command of English syntax. In contrast, novice teachers focused predominantly on either analysing language features or comprehending the ideas communicated in the text. In addition, the participating expert teachers were also observed to rate consistently lower various aspects of sample compositions than the novice teachers did.

Apart from studies that explored the effect of raters' general English teaching experience, several researchers have observed an influence of raters' experience with the culture and language of ESL writers on L2 writing evaluation. For example, Hamp-Lyons (1989) noted that native English speakers may become either positively or negatively biased toward ESL writing based on their experience with the culture and language of the writers. Similarly, Land and Whitley (1989) have reported that readers with bilingual or multilingual

experience may value different writing styles. Other researchers have observed that faculty members with more exposure to ESL students may be either more tolerant of their language errors (Vann, Lorenz and Meyer, 1991) or more lenient in the holistic evaluation of ESL essays (Song and Caruso, 1996). Together, these observations suggest that the amount of exposure to the language and culture of ESL students may affect raters' judgments.

As the above review shows, previous studies have defined teaching experience in terms of either general English teaching experience (Cumming, 1990; Hughes and Lascaratou, 1982) or exposure to L2 culture and language (Hamp-Lyons, 1989; Land and Whitley, 1989; Song and Caruso, 1996; Vann *et al*., 1991). Since experience in teaching L2 writing is an important factor directly influencing teachers' evaluation of L2 writing, research needs to isolate teaching experience in L2 writing as a principal variable of investigation. In view of this need, the present study, as part of a larger study on how teachers evaluate Chinese students' English writing, aims to identify whether the same text feature or piece of writing may evoke different responses from teacher-raters with varying years of experience teaching L2 writing. Such investigation, together with previous research on teacher-raters' general English teaching background and exposure to the L2 culture and language, is a prerequisite for improving the validity of criteria and procedures in writing evaluation (Connor-Linton, 1995a), a way to trace differences in the teaching beliefs and practices of various teacher-raters (Connor-Linton, 1995b), and ultimately a resource to inform teachers how L2 writing instruction may help students develop a sense of audience (Hamp-Lyons, 1989). The following question was formulated to focus the present study:

> What is the relationship between years of teaching L2 writing and raters' holistic scores and qualitative comments?

Method

Teacher-raters

A total of 46 teacher-raters from 23 tertiary institutes in 12 cities in China participated in the study. As most of the English writing programs in Chinese universities are taught jointly by local and expatriate teachers, we recruited an equal number of volunteers from each group. The 23 expatriate teachers, all native English speakers, responded to an invitation sent to a list of 70 teachers with the help of a Christian organization that assisted Chinese universities in hiring native English teachers. The 23 local Chinese teachers were volunteers from the participating university. They were mostly in-service teacher-trainees from various tertiary institutions. All participating teachers completed a questionnaire that requested demographic information including age, native language, general English teaching experience, and teaching experience in ESL or EFL

(English as a Foreign Language) writing. Of these variables, experience teaching ESL/EFL writing was the independent variable of interest. We included teaching experiences in both ESL and EFL contexts as some of the native English teachers might have taught ESL writing in their home countries before teaching EFL writing in China.

Table 1 summarizes the participants' years of teaching ESL/EFL writing in relation to their L1. The number of years of teaching ESL/EFL writing was aggregated into three groups (0 years, 1–4 years and ≥5 years) based on the distribution of the data. Such grouping may help identify how these teachers vary in evaluating students' essays as they gain experience at different stages of their teaching careers. As illustrated in Table 1, seven of the eight participants who had no experience in teaching English writing were Chinese, whereas twelve of thirteen participants who had taught English writing for five years or more were expatriate teachers. This confirms our observation that primarily English native speakers are hired by universities in China to teach writing, a language skill that many Chinese teachers might feel less confident teaching. Although they had less experience in teaching English writing, all Chinese participants had taught general English. Like most in-service teacher-trainees in China, many Chinese participants were experienced English teachers without a graduate degree.

Table 1: Teacher profiles

	Years of teaching ESL/EFL writing				
Country of origin	0 years	1–4 years	≥5 years	Total	%
China	7	15	1	23	50.0
United States	0	6	8	14	30.4
Britain	1	4	0	5	10.9
Canada	0	0	2	2	4.3
Norway	0	0	2	2	4.3
Total	8	25	13	46	100

Evaluation of students' essays

Ten essays were randomly selected from 86 in-class writing assignments. (Every eighth essay was selected from the whole set, collected in no particular order. See Appendix A for four sample texts.) Three teachers administered the writing task in their 50-minute writing classes for third-year English majors in a large Chinese university. At the time of data collection, the students were practising how to write argumentative essays as part of their academic writing program. The writing prompt was suggested by the three classroom teachers to fit the task into their teaching routines:

> Nowadays with the popularity of television people gain daily news more conveniently. Some people even begin to play down the advantage of newspapers arguing that it is time that they were replaced by television. To what extent do you agree or disagree with this statement? Give support for your argument.

The ten essays were sent to the participating teachers who evaluated them using a 10-point scale and provided three comments or reasons justifying the scores. No evaluation criteria for the 10-point scale were provided, so that these teachers had only their own experience to guide them in their quantitative and qualitative evaluations. We were aware that, without rating criteria to guide essay raters, it would be difficult to interpret what a particular score meant when given by different raters. However, previous research, such as Cumming (1990), did not use any criteria or analytical categories for their rating scales in order to find out how raters defined the criteria themselves. Based on the assumption that a non-specified scoring procedure might be more sensitive to evaluation differences associated with the teaching experience of raters, teachers in the present study were asked to observe the following instructions as they evaluated the ten essays:

> This project aims to find out how teachers of English rate university students' essays. Please read and rate the 10 essays provided using a 10-point scale (10 points being the highest on the scale) and then state, by the order of importance, three reasons or characteristics in each essay that you think have most influenced your rating of that essay (The first reason being the most important).

Coding of qualitative comments

A coding scheme was developed to compare teachers' comments. Based on our initial observations, we found that the teachers' comments each typically contained an adjective such as "good/strong" or "poor/weak," indicating whether the comment was positive or negative, and a content word indicating a general or particular textual feature such as "general quality," content," "organization," "language" and "length" that the teachers chose to focus on. Based on these key words, we coded the comments as positive or negative in five major categories: comments for the *general quality*, *content*, *organization*, *language* and *length* of the essays. The categories of *content*, *organization* and *language* were then each further analysed to identify subcategories of *general* and various *specific* comments. The subcategories of specific comments were *ideas* or *arguments* under the category of *content*; *paragraph* and *transitions* under the category of *organization*; and *intelligibility*, *accuracy*, and *fluency* under the category of *language*. Thus, a total of twelve categories were generated. (See Appendix B for definitions and examples of each category.) To check intercoder

reliability, two of us independently coded comments from ten teachers and reached an agreement of 95 percent. Based on the coding of the entire data set, we identified a total of 1,299 comments, both positive and negative, in terms of the twelve categories indicating how these teachers perceived the quality of the ten student essays. Some teachers gave fewer than three comments for some of the essays.

Data analysis

The holistic scores and qualitative judgments for the ten essays were analysed to determine to what degree years of teaching ESL/EFL writing was a significant factor in teachers' evaluations. To compare the scores, we first ran reliability tests to determine the extent to which each group of teachers, defined by years of teaching ESL/EFL writing, agreed on their holistic scores for the ten essays. Then ANOVA and post hoc Tukey tests were performed to assess differences in the mean scores given for each essay by various groups of teachers. For essays that differed in the holistic scores, we then ran ANOVA and post hoc Tukey tests to compare the mean frequencies of the 24 comments (12 categories of positive and 12 categories of negative comments) to determine whether any differences in these teachers' qualitative judgments might explain the differences in scores.

Results

Reliability

Reliability of teacher groups defined according to varying years of teaching ESL/EFL writing was computed based on intraclass correlation coefficients. Reporting the correlations, Table 2 shows that the reliability coefficients ranged from .68 to .84, suggesting different levels of agreement of each group as a whole in responding to the set of ten essays.

Table 2: Comparison of Reliability coefficients of the teacher-raters with varying years of teaching experience (Alpha)

0 years (n = 8)	1–4 years (n = 25)	$\geq$5 years (n = 13)
.68	.84	.75

The group of teachers who had no experience in teaching ESL/EFL writing achieved the lowest reliability (Alpha = .68).[1] In comparison, the most consistent teachers were those who had taught ESL/EFL writing for one to four years (Alpha = .84). The most experienced group (five or more years), though more consistent (Alpha = .75) than the inexperienced writing teachers, were less consistent than teachers who had taught writing for one to four years. It seems that some experience in teaching ESL/EFL writing (one to four years)

helped teachers to be more consistent in evaluating students' essays. What is not easily explainable, however, is that the teachers with the most experience (five or more years of teaching ESL/EFL writing) did not achieve the highest reliability.

Evaluation scores

Table 3 summarizes the means and standard deviations of the scores of teacher groups defined by different years of teaching ESL/EFL writing.[2] As the table shows, the mean scores of the ten essays ranged from a low of 5.45 (Essay 6) to a high of 8.01 (Essay 9). The group means suggest a tendency for teachers with no experience in teaching ESL/EFL writing to give the highest scores, followed by the more and the most experienced groups (Group means of 7.26, 6.97 and 6.05). This tendency for the least experienced groups to give higher scores was supported by one-way ANOVA and post hoc Tukey tests ($p < .05$) that indicated significant group differences in the scores for four of the ten essays (see Appendix C for statistical details). As Table 3 illustrates, teachers who had no experience in teaching ESL/EFL writing gave higher scores than the most experienced writing teachers did for Essays 1, 2, 4 and 10. Similar patterns were found between teachers who had more experience in teaching ESL/EFL writing and the most experienced writing teachers; the former gave significantly higher scores than the latter for Essays 2, 4 and 10.

Qualitative comments for essays suggesting differences in scores

We compared the mean frequencies of the twelve categories of comments for Essays 1, 2, 4 and 10 that showed differences in scores to determine whether teachers also differed in their evaluative comments, which might explain the differences in scores. ANOVA and Tukey tests revealed that each of the four essays received at least one differing qualitative judgment among various teacher groups ($p < .05$). Depending on the individual essays, significant differences were found between the inexperienced and more and/or most experienced writing teachers on qualities such as *general organization*, *language fluency*, *ideas* and *general language use*.

For Essay 1, the inexperienced writing teachers were more supportive than were the more experienced and the most experienced writing teachers for the *general organization* of the essay (means of 0.50 vs. 0.12 and 0.00, F = 6.00, $p < .01$). None of the other positive or negative comments showed any significant differences among the three groups. The essay (see Appendix A) starts with an introduction of the topic: "Since the invention of TV, newspaper business has deciding [declined?]. Fewer and fewer people are reading newspaper and someone even declares that newspaper is dying out soon." It then comments on the advantages and disadvantages of the images and sound

Table 3: Teachers' experience of teaching ESL/EFL writing and their mean holistic scores of the ten essays

	Years of teaching ESL/EFL writing				
Essays	0 years (n = 8)	1–4 years (n = 23)	≥5 years (n = 13)	F	df
1	7.81 (1.07)	6.96 (1.54)	5.72 (2.38)	3.90*	2
2	7.25 (0.76)	7.04 (1.29)	5.77 (0.83)	6.78*	2
3	7.44 (0.94)	7.38 (1.33)	6.54 (1.51)	1.93	2
4	6.38 (0.95)	5.98 (1.66)	4.41 (1.62)	5.58*	2
5	7.94 (0.68)	7.64 (1.20)	7.18 (1.30)	1.18	2
6	5.88 (1.25)	5.60 (1.23)	4.92 (1.88)	1.37	2
7	7.90 (0.96)	7.26 (1.94)	6.87 (2.16)	0.74	2
8	7.50 (1.07)	7.11 (1.09)	6.43 (1.58)	2.11	2
9	7.81 (0.80)	8.15 (1.08)	7.85 (1.45)	0.43	2
10	6.69 (1.19)	6.62 (1.65)	4.79 (2.20)	5.12*	2
Group M	7.26 (0.97)	6.97 (1.40)	6.05 (1.70)		

Notes:
*$p < .05$.
Means that differ significantly according to Tukey's HSD are indicated by arrows and joined by a line.
Standard deviations are given in brackets.

provided by television and concludes with the statement that television is not bad if we make good use of it. This inductive organization of the essay is different from the conventional English essay development, which typically starts with a thesis statement followed by supporting arguments. The less positive comments or attitudes of the more and most experienced teachers suggest that they were probably more sensitive to such violations of English expository conventions compared with the inexperienced teachers. Recalling that the inexperienced writing teachers gave significantly higher scores for Essay 1, their positive comments for *general organization* might have been a reason for their higher scores. Alternatively, the lower scores from the most experienced writing teachers might be traced to their less positive attitudes based on the same criterion.

For Essay 2, there is a significant difference in the negative comments for *language fluency* between the inexperienced and most experienced teacher groups. As noted earlier, the inexperienced writing teachers gave significantly higher scores for the essay. They actually supported their higher scores for *fluency* with fewer negative comments than the most experienced teachers did (mean of 0.13 vs. 0.69, F = 6.34, $p < .01$). An initial reading of Essay 2 (see Appendix A) shows that the student used similar syntax throughout. Most of the sentences (14 out of 22) contain the structure of "I/We/It can/will/want . . . ." The repetition of such syntax, a language problem hindering smooth reading of the essay, seemed to be more salient for the experienced teachers. The different attitudes toward the quality of *fluency* in Essay 2 suggest the influence of teaching experience in assessing L2 writing.

For Essay 4, the only significant difference pertained to positive comments for *ideas*. The inexperienced writing teachers gave more positive comments for *ideas* than the more and the most experienced groups did (mean of 0.63 vs. 0.16 and 0.15, F = 4.48, $p < .01$). In terms of *ideas*, Essay 4 (see Appendix A) argues that newspapers are better than television because the former offers detailed written information that allows readers to make critical reflections. The writer's preference for written communication, which is different from the general public's preference for the sound and images provided by television communication, seems to have won positive comments from the less experienced writing teachers. It is difficult, though, to explain why those more experienced writing teachers did not value such individual or personal ideas in their evaluations as much. Like the qualities of *general organization* of Essay 1 and *fluency* of Essay 2, the quality of *ideas* of Essay 4 revealed how teachers with diverse experience in teaching ESL/EFL writing responded to a particular student's text.

For Essay 10, the inexperienced writing teachers were found to be the only group that gave positive comments for *general language* quality (mean frequencies of 0.25 vs. 0.00 and 0.00, F = 7.71, $p < .01$). We examined Essay 10

(see Appendix A) and found various language problems such as the following (problematic parts are underlined):

1. These are quite advanced than those by newspaper.
2. Newspaper is essential as well as television.
3. If it be replaced, the world become dim a half.
4. Television play a great role in our life, exactly.
5. . . . my assignments are seldom finished in a good way for periods.
6. Television shows us and may be make some comments, while newspaper offer another expanding fields for the public to express themselves –thus, they make the hits.

The fact that inexperienced writing teachers gave more positive comments for the language quality despite the errors illustrated above suggests a more lenient evaluation of language from these teachers compared with the more experienced writing teachers. Since seven out of eight of the inexperienced writing teachers were local Chinese teachers, this leniency might have reflected, apart from a lack of teaching experience, the particular language standard used by the participating non-native English-speaking teachers.

Discussion

Based on the comparisons of the mean scores for individual essays in association with teachers' experience, the present study suggests that the most experienced ESL/EFL writing teachers gave much lower scores than the less or least experienced teachers did for essays 1, 2, 4 and 10. This finding contributes to our understanding pertaining to the influence of teaching background on L2 writing evaluation. On the one hand, it echoes Cumming's (1990) observation that the novice teachers in his study consistently rated various aspects of sample compositions higher than the experienced teachers did. The present study suggests that like experience in teaching general English, experience in teaching ESL/EFL writing seems to make teacher-raters stricter in their writing evaluations. On the other hand, the present finding that the L2 writing instructors were less lenient in their evaluation as they became more experienced or more aware of students' weaknesses differs from Song and Caruso's (1996) observation that faculty members who had had previous experience with ESL students became more sympathetic to ESL students' writing problems and therefore gave more lenient holistic evaluations. The different finding in the present study suggests that the experience of teaching L2 writing is different or more than just the exposure to L2 students and their writing that many faculty members experience. Different readings of students' texts have been a major concern of scholars with respect to the fairness in L2 writing evaluation (Connor-Linton,

1995b; Silva, 1997). The present study suggests that using raters with similar teaching experience might alleviate some of the potential differences.

Analyses of the qualitative comments for Essays 1, 2, 4 and 10 reveal differences between various teacher groups in their qualitative comments for the four essays. As we illustrated earlier, the most experienced writing teachers gave much lower scores when they evaluated either less positively or more negatively features pertaining to the *general organization*, *ideas*, *general language* and *fluency*. The inexperienced writing teachers, who were more positive or less negative in regard to these features, consequently gave higher scores for the same essays. The fact that the most experienced ESL/EFL writing teachers were stricter or less positive regarding these aspects of the essays implies that certain evaluative attitudes were associated with the teaching experience of the raters. The present study, having isolated years of teaching ESL/EFL writing as the main focus of research, contributes to our understanding of how experience in teaching L2 writing might help predict and explain some reader responses to student writing. However, since seven of the eight teachers with no experience teaching ESL/EFL writing were Chinese, other factors might have influenced their rating such as more familiarity with Chinese students' English or their own English proficiency (Shi, 2001).

With findings of differences in the holistic and qualitative evaluations between the most experienced ESL/EFL writing teachers, the present study also implies similarities among participants. The fact that significant differences were found in only four of the ten essays suggests that experienced and inexperienced writing teachers shared evaluations on the other essays. The similarities and differences among the participating teachers remind us of variables other than teaching experience in L2 writing, such as L1, age, educational background, and academic status, that are important factors in the English-teaching context in China. In other words, though having varying experiences in teaching writing, these teacher groups might share or differ in other background variables which might have contributed to the similarities and differences in their evaluations. The present study represents a promising pilot study for a more thorough research project with far more careful control and management of the variables to delineate meaningful patterns in a real-life teaching context such as China.

Conclusion

The present study investigated the nature and manner in which teachers with diverse experience in teaching ESL/EFL writing differed in their quantitative and qualitative judgments when evaluating the same set of student essays. Specifically, it examined how 46 teachers evaluated ten essays by comparing their self-generated evaluation criteria to the number of years they had taught

ESL/EFL writing. Results show that the most experienced ESL/EFL writing teachers gave much lower scores than did the less or least experienced teachers for Essays 1, 2, 4 and 10. In justifying their lower scores for these essays, the experienced ESL/EFL writing teachers gave either fewer positive or more negative comments for *general organization*, *language fluency*, *ideas* and *general language*.

Although the present study helps us understand the impact of teaching experience in regard to L2 writing evaluation, it has at least two limitations. First, although it is necessary to focus on a variable in the real-life context to identify how the profession works in a given setting such as China, we are aware that the results of the study may be limited when such a crucial variable escapes a certain degree of control. For example, we sampled both native English- and non-native English-speaking teacher-raters to represent the two groups of writing instructors in Chinese universities, locals and expatriates. However, difference in raters' L1 backgrounds is also a variable influencing the way raters assessed L2 writing (see for example, Shi, 2001). Since most of the least experienced teachers in this study were local teachers while the most experienced teachers were expatriates, differences in evaluation between these two groups might result not only from their diverse teaching experience but also from their differing L1 backgrounds. Future research, therefore, should control for this variable by sampling participants in either group. Second, the present study lacks attention to individual variability among teachers. As Vaughan (1991) put it, individual teacher-raters may "focus on different essay elements and perhaps have individual approaches to reading essays" (p. 120). Future studies using a more purposeful sampling method or focusing on individual cases might zero in on how teaching experience has an impact on individual teacher-raters' judgments in relation to their personal and cultural backgrounds. In sum, the present study only portrays a small portion of what is certainly a much larger, more variable and more complex situation. Preliminary as it is, we hope it shows the way to a number of exciting research possibilities on L2 writing evaluation.

Notes

This project was supported by a Humanities and Social Sciences Research Grant from the University of British Columbia granted to the first author. We thank the participating students and teachers, as well as Joe Belanger, Stephen Carey, Marcel Sauvé and Monique Bournot-Trites, and three anonymous reviewers and the English editor of the *CJAL* for their comments on earlier drafts of the paper.

[1] Alpha, a model to assess internal consistency, represents average inter-item correlation. It is appropriate for our study compared with Pearson or Spearman, which are bivariate coefficients and check consistency for only two items at a time.

[2] The format for this table was inspired by Slabakova (1999), Table 2, p. 298.

References

Connor-Linton, J. 1995a. "Looking behind the curtain: What do L2 composition ratings really mean?" *TESOL Quarterly*, 29, pp. 762–765.

Connor-Linton, J. 1995b. "Crosscultural comparison of writing standards: American ESL and Japanese EFL." *World Englishes*, 14, pp. 99–115.

Cumming, A. 1990. "Expertise in evaluating second language compositions." *Language Testing*, 7, pp. 31–51.

Hamp-Lyons, L. 1989. "Raters respond to rhetoric in writing." In *Interlingual processes*. H.W. Dechert and M. Raupach (eds.). Tübingen: Gunter Narr Verlag, pp. 229–244.

Hughes, A. and C. Lascaratou. 1982. "Competing criteria for error gravity." *ELT Journal*, 36, pp. 175–182.

Land, R. and C. Whitley. 1989. "Evaluating second language essays in regular composition classes: Toward a pluralistic U.S. rhetoric." In *Richness in Writing*. D. Johnson and D. Roen (eds.). New York: Longman, pp.289–293.

Shi, L. 2001. "Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing." *Language Testing*, 18, pp. 303–325.

Silva, T. 1997. "On the ethical treatment of ESL writers." *TESOL Quarterly*, 21, pp. 359–363.

Slabakova, R. 1999. "The parameter of aspect in second language acquisition." *Second Language Research*, 15, pp. 283–317.

Song, B. and L. Caruso. 1996. "Do English and ESL faculty differ in evaluating the essays of native English-speaking and ESL students?" *Journal of Second Language Writing*, 5, pp. 163–82.

Vann, R.J., F.O. Lorenz and D.M. Meyer. 1991. "Error Gravity: Faculty response to errors in the written discourse of nonnative speakers of English." In *Assessing second language writing in academic contexts*. L. Hamp-Lyons (ed.). Norwood, NJ: Ablex, pp. 181–193.

Vaughan, C. 1991. "Holistic assessment: What goes on in the rater's mind?" In *Assessing second language writing in academic contexts*. L. Hamp-Lyons (ed.). Norwood, NJ: Ablex, pp. 111–125.

Appendix A:
Writing samples that had different evaluations from teachers

Note: Errors and mistakes are retained verbatim from the original students' essays.

Essay 1.

Since the invention of TV, newspaper business has deciding. Fewer and fewer people are reading newspaper and someone even declares that newspaper is dying out soon.

Judging from the means of providing information, TV has its unprevailable advantage over the newspaper. It provides visions and sounds. It was a great event that the first day people could see with their eyes what had happened in the other part of the world. Thus the information on TV become powerful for pictures are different from printed words—seeing is believing. In this sense, TV can present the news, the events more vividly and carry more information from several angles. It should be a better source.

However, the truth is that newspaper is more reliable simply because of the commercialization of TV. TV is powerful and people abuse its power. Most of what we get from TV is not information, but a kind of entertainment. The news reported on TV are usually in terse, popular language with impressive images, but without much profounding critics. We find more good critics on newspapers. The TV people don't give critical opinions, they are busy making questionaires about masses' taste and fussy with soapy shows. And if they have an attitude, (I feel happy as well as worried about it.) they can use, manipulate their powerful weapon—TV without letting us know. Pictures clipped, words omitted, repetitions on certain facts, all these can give us a totally wrong idea. Yet, the worst thing is we believe the news, for seeing is believing. One Hong Kong banker was interviewed by foreign journalists about his opinion on Hong Kong's future. He said "There is difficulty ahead, but I have the full confidence." On screen , you just hear "There is difficulty ahead" and see a shot of his lowered head. What is the truth? I remember one film named "making city" starred by Dustin Huffman. The protagonist doesn't tend to harm anyone, but is made the image as a terrorist by the media and shoots himself at last. When Dustin cries, "We killed him" in the end of the film, I feel there's something much worse than entertainment that TV can provide—the false.

Of course, TV itself is not bad. It's like money and depends on how we use it. The problem with the masses is that we are easily taken in by what we see and by indulging too much in its entertainment, we don't think and become slave of it. In this sense, the "ancient" means of data leaves us more space for thinking.

Essay 2.

Which is a better source of information: Newspapers or Television? In my opinion, I think that television is a better source of information. Maybe it

is related to that I like watching television more than reading newspapers. Because on television, we can at the same time, receive audio and visual information. For example, if we watch news program on TV, we can hear that where the event happened, when and why the event happened. We can also watch the scene of the event, so as to be more impressed about the event. In this way, we will not hear the statements of announcers, but see the spot scene. I think by watching television, we can learn more than reading newspapers. We can be more impressed and informed, and learn more information about many great events in the current world. And the other advantage of television is that, we can be informed more quickly, more actively, more accurately and more animately. And these advantages are not possessed by newspapers.

Now I can give you a typical example, about the latest World Cup in Paris, we audiences want to watch the violent matches, especially the playoffs. We don't only want the results of the matches, which wins and which fails. We also want to experience the matches with the players. We are eager to appreciate the match. These, newspapers can not supply. It can, to the most, describe the match animately but can not give us the scene, the image! So we audiences are not satisfied.

We don't deny that newspapers can be a good source of information. It can give us a lot. But, generally speaking, television is a better source of information. That's a obvious matter of fact. It's unsuspectable!

Essay 4.

I don't agree that the newspaper will be replaced by TV. The fact is that TV has come into being for about 40 years, newspaper still exists and some even expand their business, just because of its advantage.

As sources of information, TV only provides us with facts, turning us into dumb creatures while newspaper makes us thinkers. News on TV flashes so quickly that we hardly have time to think critically about an event, such as the Financial Crisis. In contrast, newspaper notes everything down, both the fact and various personal opinions. It helps to form our own idea. Thus we are the real human rather than the tool.

Newspaper also can fully reveal the witchery of words, subtle articles as fiction, lyric and even Shakespear's drama can be seen on it. They give us a lot of fun.

Thus newspaper will become dominant information source for people, especially those high qualified, and it will never disappear.

Essay 10.

Though television is hotly popular nowadays, newspaper is still expressing itself attractively instead of being replaced.

Our society is a most colorful one. People live in it and feel it through various ways among which television and newspaper are the two most essential ones.

Television shows people what's going on in a magnificant specific way as well as by its fastest correspondences. People feel the atmosphere of the events of spot just at home, which is convient and economic, isn't it? These are quite advanced than those by newspaper. Nevertheless, newspaper also has its dear merits. It is easy to carry with. Mostly, wherever you go, you can fatch a newspaper from any newspaper vendors since they are nowadays wandering through every corner of the world. And it is quite flexible for you to choose what you are interested in to read through those obvious captions in it, therefore, you save time and energe, compared with the way when you see television you have to follow its program schedule.

Television play a great role in our life, exactly. As for my family, we are totally captured by it, meanwhile, it cause problems. Sometimes, I am not willing to do anything else, therefore, my assignments are seldom finished in a good way for periods. At many moments, I and my brother are in a war for controlling the TV to see the programs we are respectively interested in. Then now, with regard to newspaper, I know, these will be more simply solved, for we can buy the newspaper on a low price and satisfy our respective tastes. In addition, through reading newspaper, you can make many pauses to do something else. I usually do my homework and for a short while read the newspaper as relaxation.

Another great point is that nowadays, television and newspaper are helping each other. As a medium, they play roles of arousing attentions or interests of people to the things that are happening. Television shows us and may be make some comments, while newspaper offer another expanding fields for the public to express themselves — thus, they make the hits.

Newspaper is essential as well as television. If it be replaced, the world become dim a half. So how can this happen?

Appendix B: Coding categories of qualitative comments

Major categories	Sub-categories	Definitions	Examples of positive/ *negative** comments
General		General comments for overall quality of writing.	– well written – *it fails to complete the task*
Content	General	General comments for content.	– good contents – *content shallow*
	Ideas	General or specific comments for ideas and thesis.	– good ideas – *poor ideas*
	Arguments	General or specific comments for aspects of arguments such as balance, use of comparison, counter-arguments, support, uses of details or examples, clarity, originality, relevance, logic, depth, objectivity, conciseness and development.	– good argument – *poor argument* – arguments balanced – *lack of arguments on the newspaper issue* – arguments well supported – *arguments not very well supported*
Organization	General	General comments for organization.	– excellent organization – *weak organization*
	Paragraphs	Comments concerning paragraphs.	– paragraphs are well arranged – *paragraphs are poorly organized*
	Transitions	Comments concerning transitions, coherence and cohesion.	– good transitions – *bad use of transition words* – coherent – *lacks coherence*
Language	General	General comments for language.	– language good – *poor English*
	Intelligibility	Comments on whether the language is clear or easy to understand.	– easy to read and follow – *meaning unclear*
	Accuracy	General comments for accuracy or specific comments for word use, grammar and mechanics.	– accurate language – *too many errors*
	Fluency	Comments for fluency, conciseness, maturity, naturalness, appropriateness and vividness of language.	– fluent language – *language not smooth* – concise language – *wordy*
Length		Comments on whether the writer has fulfilled the word limit.	– about 250 words – *too short*

*Negative comments are in italics.

Appendix C:
ANOVA tests of the scores for ten essays by three groups of teachers

Essay		SS	df	MS	F	*p*
1	Between Groups	24.047	2	12.024	3.904	.028
	Within Groups	132.446	43	3.080		
2	Between Groups	16.494	2	8.247	6.777	.003
	Within Groups	52.325	43	1.217		
3	Between Groups	6.845	2	3.423	1.934	.157
	Within Groups	76.090	43	1.770		
4	Between Groups	26.895	2	13.447	5.579	.007
	Within Groups	103.644	43	2.410		
5	Between Groups	3.192	2	1.596	1.182	.317
	Within Groups	56.698	43	1.350		
6	Between Groups	5.722	2	2.861	1.374	.264
	Within Groups	89.552	43	2.083		
7	Between Groups	5.262	2	2.631	0.743	.482
	Within Groups	148.624	43	3.539		
8	Between Groups	6.518	2	3.259	2.114	.133
	Within Groups	66.294	43	1.542		
9	Between Groups	1.165	2	0.582	0.434	.651
	Within Groups	57.703	43	1.342		
10	Between Groups	31.741	2	15.871	5.123	.010
	Within Groups	133.218	43	3.098		