Effects of Self-assessment Training on Chinese Students' Performance on College English Writing Tests

Huiqing Zheng, Jianbin Huang, and Ying Chen

Abstract

In this empirical study of students' self-assessment in College English writing tests, 189 freshman and sophomore students in eight classes were taught over an 8-week period how to assess their writing work. Both quantitative and qualitative approaches were employed. Specifically, two kinds of relationship were analyzed: first, the relationship between the rating criteria instruction and the improvement of student self-assessment and second, the relationship between the rating criteria instruction and college English writing performance. The results showed statistical significance as follows. Firstly, students could perform self-assessment in writing reasonably well. Secondly, the instruction of the scoring rubric contributed to the improvement of self-assessment in writing and the overall improvement was significant. Thirdly, students' overall composition performance was enhanced. Such findings point to the positive influence of self-assessment on the learning process, providing it is applied meaningfully and taught well.

Key terms: self-assessment; instruction of rating criteria; CET writing performance; positive effects

1. Introduction

1.1 Research Background

During the last 30 years there has been a surge of interest in methods for self-assessment of foreign language proficiency. According to Ross *et al.* (1999), few studies have examined the effects of teaching students how to self-evaluate in classroom settings over a sustained period (i.e., 4 weeks or more); seven studies viewed by Hillocks (1986), in which students were given scales for judging writing samples, reported positive effects on student performance. Similar results have been reported by Oscarson (1978, as cited in Oscarson, 1997), LeBlanc and Painchaud (1985, as cited in Oscarson, 1989), Janssen-van Dieten (1989), Arter, Spandel, Culham, & Pollard (1994, April).

Oscarson (1978, cited in Oscarson, 1997) found that adult learners studying EFL were indeed able to make fairly accurate appraisals of their own linguistic ability using a variety of scaled descriptions of performance as rating instruments.

A study of self-assessment in the language area was conducted by Arter *et al.* (1994), which gave grade 5 students direct instructions on the meaning of six traits, or ideas of essay writing. Students scored a sample of essays and applied trait analysis to their own writing over a 5-month period. The treatment group outperformed controls significantly on one of the six traits. Ross *et al.* released "Effects of Self-Evaluation Training on Narrative Writing" in 1999. In their study, 148 students in 15 grade 4-6 classrooms were taught over an 8-week period how to evaluate their work. Treatment group students became more accurate than controls in their self-evaluations. Treatment students also outperformed controls on narrative writing, but the overall effect was small.

In the present study, we attempted to apply critically the previous research findings in the self-assessment for College English Test (CET) writing, aiming to explore two kinds of relationship: first, the relationship between the rating criteria instruction and the improvement of student self-assessment; and second, the relationship between the rating criteria instruction and college English writing performance.

1.2 A Framework for the Study

Based on Athanasou's study of self-evaluations in adult education and training (Athanasou, 2005), and Shaw's framework for conceptualizing writing test performance (Shaw & Weir, 2007, p. 4), we initiated a self-assessment model for CET writing instruction

for Chinese tertiary EFL learners and educators, as shown in Figure 1.

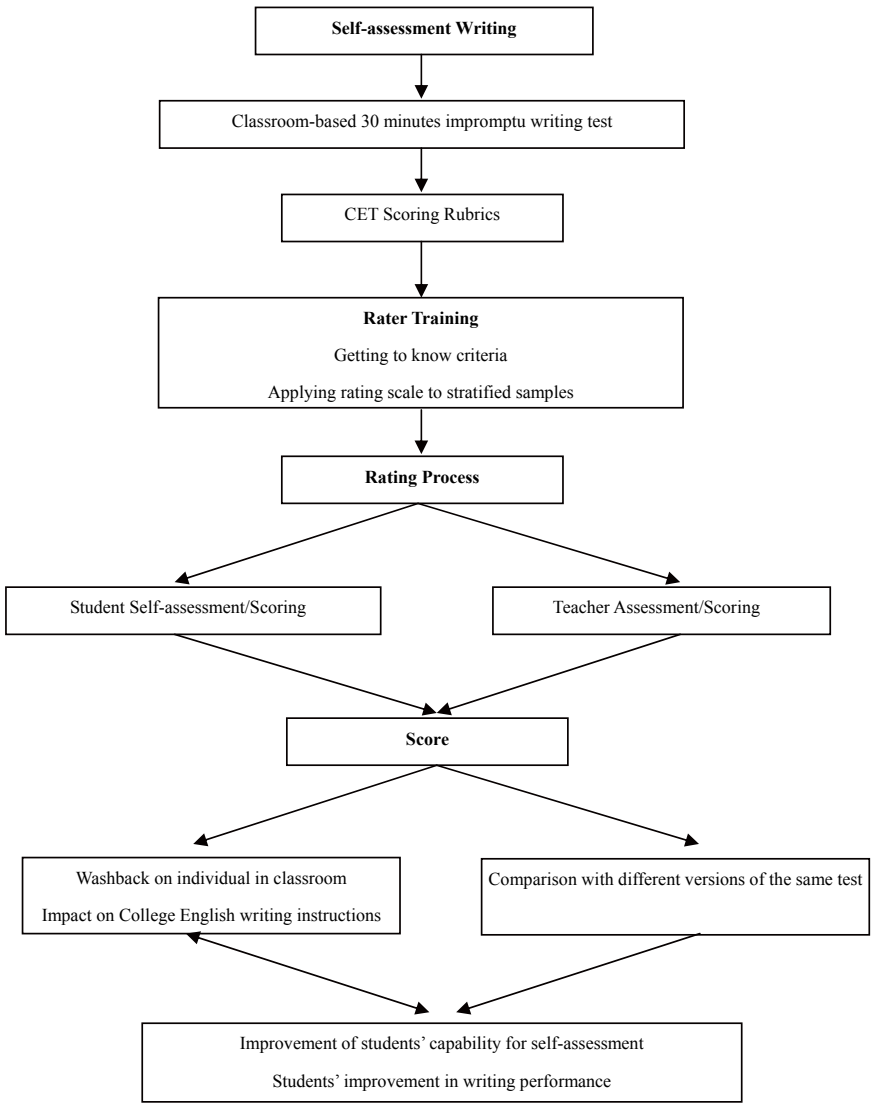

Figure 1. The Self-assessment Training Model for CET Writing Performance. In this figure, the arrows indicate the principal direction(s) of both the procedures of self-assessment and the relationships showing what has an effect on what. The timeline runs from top to bottom from the writing test administration and finally to what happens after the test event.

1.3 Research Objectives

This study sets out to examine the contribution that instruction for writing assessment criteria may offer learners in university-level

foreign language classrooms, with a particular focus on self-assessment in writing. The emphasis is first on the link between the rating criteria instruction and the improvement of student self-assessment; and second, on the relationship between the rating criteria and college English writing performance. Thus the aims of the research are to answer the following questions:

1) How accurate are college students at assessing their own essays?

2) Will rating training increase the accuracy of students' self-assessments? (We anticipate that students receiving such training would begin to evaluate their own work more accurately.)

3) Will self-assessment training contribute to language achievement? (We anticipate that focusing students' and instructors' attention to performance criteria would result in an overall enhancement of students' achievement in English writing.)

2. Methods

2.1 Participants

This research was conducted with the cooperation of 189 non-English major students at Zhejiang University, who took the course 'College English' for credits. They ranged in age from nineteen to twenty-one. They had been learning English in China for about seven or eight years respectively.

Two instructors from Zhejiang University participated in the study. One of them has an MS in applied linguistics at Zhejiang University. She had been teaching College English for three years and had 3-year experience rating the written section of the College English Test at Hangzhou CET Rating Center. The other instructor had a BA in English literature at that time. She had been teaching College English for 15 years and had likewise 10-year experience rating the written section of the College English Test at Hangzhou CET Rating Center.

2.2 Materials

Four types of materials were employed in the experiment. They are as follows:

A. Three writing test sheets each with three writing topics. The first two topics came from the writing part of both the CET Band 4 and the CET Band 6 that were used in China in January, 2005. The third writing test was made by the instructors and researchers themselves. In developing this task, attention was paid to context validity (Shaw & Weir, 2007).

B. Writing assessment criteria. A Chinese version of the Scoring Rubrics for CET4-6 was used, i.e. scoring guide for the CET 15-point rating scale.

C. Two sets of range-finders. These were exemplar scripts provided by Hangzhou CET Rating Center in January 2005 for the first two writing topics mentioned above.

D. Two sets of stratified exemplars. Two stratified samples of 11 compositions each on the first two topics (mentioned above in "A" from Hangzhou CET Rating Center in January, 2005). They were drawn from national or district sample scripts.

2.3 Experimental Procedures

Based on the authors' framework for the study of effects of self-assessment on College English writing, as shown in Figure 1, the students in the experimental group received the extra self-assessment training throughout the 8-week College English classes. Rather than being presented in a separate lesson, the scoring criteria instructions were incorporated into the regular classroom learning activities. At times, the focus on the scoring criteria instruction was explicit in that the instructors provided self-assessment training. The learners finished writing a timed essay and were asked to immediately self-rate their own work using the 15-point scale before they received a scoring criteria instruction for the writing assessment. With the acknowledgement of the scoring criteria, they were asked to self-rate the essay again. The process was repeated until they finished the 3rd timed essay and its subsequent self-rating exercise.

The in-class training activities consisted of five stages in which the instructors demonstrated a particular self-assessment technique and engaged the students in discussing their self-assessments. The instructors would explain what each criterion meant, and illustrate the levels with the set of range-finders. Then students were instructed to model the application of the criteria to the set of stratified exemplars. Finally, the students were asked to practice applying the criteria to their own writings.

The instructors or raters and students together would create a total of eleven variables, i.e. scores for the three writing tasks, not including "mean scores of two raters". The analyses of the data would account for the effects of self-assessment training on the student's college English writing.

The five stages of the empirical study are as follows:

Stage One (Week 1)

The students were first asked to write in class a 120-word letter on Writing Task 1 (Topic 1) within 30 minutes. As soon as the time expired, the students were asked to self-evaluate their own work on a holistic 1—15 rating scale on the spot. The students' score for Writing Task 1 is described as W1SF1. Five minutes later, all the writings were collected by the instructors.

Before the scoring session, the two instructors had one hour of training on rating using the Scoring Guide for CET 15-Point Rating Scales. They rated the first set of 11 stratified sample compositions independently and were provided with opportunities for the raters to see and discuss the official scores given to essays in past examinations. The first composition rating lasted nine hours with both raters under the same roof. The papers were marked in random order with student names concealed. After all the 189 papers had been marked, the two raters resolved discrepancies of two grades or above in their assessments through discussion. We refer to the first instructors' rating as W1TF1 and the second, W1TF2.

Stage Two (Week 2)

A total number of 189 students in four classes were given one and half hours of rating training using writing rating rubrics and a 5-level anchor paper. All the students completed this within a 2-day period with some students on the first day, and some on the second depending on their College English curriculum time.

Within this training, they were first asked in class to read a Chinese version of the Scoring Rubrics, i.e. Scoring Guide for CET 15-Point Rating Scales and Range-Finders (Set 1). After that, they were assigned to rate the first set of eleven sample essays, i.e. the stratified exemplars. When this was done, they were allowed to have the official scores given to essays in the past examinations and had a discussion. Finally, each participant was asked to evaluate his or her own composition again within five minutes. We refer to this score from Stage Two as W1SF2. Before they left, they were informed of W1TF (the average scores of their writing or mean scores of two raters from Stage 1). In this way they got to know how to apply the criteria to rate their own writing.

Stage Three (Week 3)

Having reviewed the Scoring Rubrics for five minutes, and been urged to try to write a better composition than the first one, the participants were assigned to finish the second writing assignment (Topic 2) within 30 minutes and to self-rate it thereafter within 5 minutes. The figure that the students gave is referred to as W2SF1.

Similarly, the second composition rating outside the classroom lasted roughly nine hours with both raters under the same roof. The papers were marked in random order with student names concealed. After all the 189 papers had been marked, the two raters resolved discrepancies of two grades or above in their assessments through discussion. We refer to the first instructor's rating as W2TF1 and the second, W2TF2.

Stage Four (Week 6)

The rating process of the second writings was similar to that of Stage Two with two differences. The instructors provided a set of range-finders and 11 stratified sample papers which corresponded to the second topic. The other change was a reduction in the time for the review of the scoring rubrics and the trial rating of the eleven stratified sample scripts was allowed for 20 minutes. The reason for the time reduction was that all the students had experienced the ratings in the previous three stages (the first session of

rating training), and the authors assumed that they had gained in rating proficiency to some extent, and therefore a quicker review of the official scores given to the eleven essays in the past examinations and a shorter discussion were desirable. Finally, each participant was asked to evaluate his or her own composition again within five minutes. We refer to this score from Stage Four as W2SF2.

Stage Five (Week 8)

The process was roughly the same as that in Stage One, but this time the students had a third writing topic, which was formulated by the instructors themselves. Besides, the students had had the writing assessment criteria already explained to them. After they completed this writing task, the students immediately self-assessed their work. The figure they gave is referred to as W3SF.

After class, outside the classroom, both the instructors rated all the scripts under the same roof at the same time. The grades both the instructors gave the students' Composition 3 were referred to as W3TF1 and W3TF2 respectively.

2.4 Data Analysis

Data obtained from the pre- and post-criteria instruction for the writing tasks were categorized in Table 1, and analyzed using SPSS (Version 14) to determine the students' improvement in both self-assessment and writing proficiency.

Table 1

Cataloging Procedures and its Corresponding Nomenclature of the Study

Tasks	Procedures	Students' rating	Rater One's rating	Rater Two's rating	Mean scores of two raters
Writing Task 1	Pre-training (Stage 1)	W1SF1	W1TF1	W1TF2	W1TF
	Post-training (Stage 2)	W1SF2			
Writing Task 2	Pre-training (Stage 3)	W2SF1	W2TF1	W2TF2	W2TF
	Post-training (Stage 4)	W2SF2			
Writing Task 3	Final Self-assessment (Stage 5)	W3SF	W3TF1	W3TF2	W3TF

To reveal how accurate the students' self-assessment could be, grade estimation errors were used.

The students' achievement in CET writing, in this study, is presented as the scores awarded by the two raters for the three writing tasks. On the grounds of expediency, the raters' mean scores were used for descriptive statistics.

A paired-samples *t*-test was used between the raters' mean scores for Writing Task 1, Writing Task 2 and Writing Task 3 to investigate whether knowing the writing scoring criteria enhances students' achievement in College English writing.

Correlating the two raters' scores for each writing task through Pearson's product-moment correlation coefficient, the result shows that there is no significant difference between the two raters' rating (r= .958, .973, .852; p=0.05). The inter-rater reliability coefficients for the two instructors were highly significant.

3. Results

3.1 Student's Accuracy for Self-assessment Pre-training

The initial focus, however, is upon the discussion of the accuracy of students' capacity for self-assessment in writing performance. Pearson's product-moment correlation coefficient was used for the first writing test scores. Correlating student and instructor grades showed a positive correlation (r= .39, p<0.05).

It is important to note that "a global correlation of this nature does not reveal a detailed picture of where the similarities and differences between students and instructors might lie"(Longhurst N. & Norton L. S, 1997). To do this more detailed analysis, grade estimation errors were used. Grade estimation errors, i.e. the variable "W1E1", were calculated by subtracting the mean score awarded by raters (W1TF) from the grade (W1SF1) students gave themselves for Writing Task 1. The variable 'W1E1" was rounded into an integer and computed. The frequency of descriptive statistics of W1E1 was used. The results are presented in Figure 2(A).

A brief look at Figure 2(A) reveals that the distribution of error scores is strongly skewed towards over-estimation. What is striking about the graph is the high level of symmetry centered around the plus 2, and its median is plus 2.

3.2 Student's Accuracy for Self-assessment Post-training

For our second question as to whether students' rating training increased the accuracy of students' self-assessment, the correlation coefficient and the grade estimation errors for the three writing tasks were investigated to find whether the correlation coefficient had increased or not, whether the grade estimation errors had been diminished or not, and, in both categories, to what extent.

The grades students gave themselves for their own writings were employed to be correlated with the mean scores the raters awarded the students for their writings. If the correlation coefficient increases each time, the increase of the accuracy of the student's self-assessment is indicated. Pearson's correlation coefficient was used again for the students' grades and instructors' ratings for all the three writing tasks.

Correlating student and instructor grades showed a positive correlation in Table 2.

Table 2

Students' Self-assessed Grades Related to Mean Scores of Raters

Grades		Correlation coefficient	Significance level
W1TF	W1SF1	0.39	0.01
	W1SF2	0.55	0.01
W2TF	W2SF1	0.46	0.01
	W2SF2	0.69	0.01
W3TF	W3SF	0.53	0.01

There is a significant increase in the accuracy of the students' self-assessment for their own writing after their rating training, where r increases from .39 to .53. The contrast between the first rating and the second rating for the same composition is striking. The correlation coefficient for Writing Task 1 had increased considerably after the first session of rating training (to .55). The greatest difference occurred after the second session of rating training, where there was an increase in the correlation coefficient for the students' second rating of Writing Task 2 (r=0.69, p<0.05), representing a one-third increase from the students' first rating of Writing Task 2 (r=0.46, p<0.05). The students' enhancement in accuracy of the self-assessment for their writings can be further confirmed by the comparison between the first rating of Writing Task 1 and their rating of Writing Task 3. After two sessions of rating training, the authors found a marked increase in the correlation coefficient (r=0.58, p<0.05) for Writing Task 3, from the original 0.39 (p<0.05) for Writing Task 1.

Comparison of the correlation coefficient between students' and instructors' grade post training shows improvements in the accuracy of students' self-assessment, thus the grade estimation errors for all the three compositions were investigated. They were calculated by subtracting the mean raw score awarded by the raters from the grade students gave themselves. As the term W1E1 refers to the students' estimation errors for their first rating of Writing Task 1 before the rating training; the term W1E2 refers to the students' estimation errors for their second rating of the same writing task, i.e. Writing Task 1 after the first session of training. The term W2E1 refers to the students' estimation errors for their first rating of Writing Task 2 upon finishing writing before the second rating training session; the term W2E2 refers to the students' estimation errors for their second rating of the same writing task, i.e. Writing Task 2 after the second session of rating training. The term W3E refers to the students' estimation errors for their rating of Writing Task 3 upon finishing writing after having had two previous sessions of rating training. A paired-samples *t*-test was used. The results are presented in Table 3.

From Table 3, within Writing Task 1, the standard error mean has decreased by 0.01. Through the paired-samples *t*-test, the result shows that there is a marked difference in the students' estimation errors between the students' first self-rating of Writing Task 1 without any rating training and their second self-rating of Writing Task 1 after the first session of rating training, t=11.83, p<0.01.

Table 3

Grade Estimation Errors for Three Writing Tasks.

Pairs		Mean Error	N	Std. Deviation	Std. Error Mean
Writing Task 1	W1E1	1.80	189	1.74	.13
	W1E2	.51	189	1.69	.12
Writing Task 2	W2E1	.20	189	1.48	.11
	W2E2	-.33	189	1.28	.09
Writing Task 3	W1E1	1.8	189	1.74	.13
	W3E	-.01	189	1.42	.10

This is also illustrated by the increase of accuracy through the trained self-rating errors by the students from Stage 2, as shown in Figure 2(B). The graph in Figure 2 (B) reveals a high level of symmetry around the plus 1, between minus 1 and plus 2; the error scores are nearly normally distributed, and its median is plus 1. Compared with Figure 2 (A), the percentage of perfect accuracy, i.e. the estimation error is zero, doubles for the second rating. Alongside this, the percentage of overestimation (two grades above the considered correct value) reduces greatly from 40% to 12 %.

Similarly, within Writing Task 2 in Table 3, the standard error mean has decreased by 0.02. Through the paired-samples *t*-test, the result shows that there is a significant difference in the students' grade estimation errors between the students' first self-rating of Writing Task 2 prior to the second session of rating training and their second self-rating of Writing Task 2 after the second session of rating training, t=6.7, p< .01.

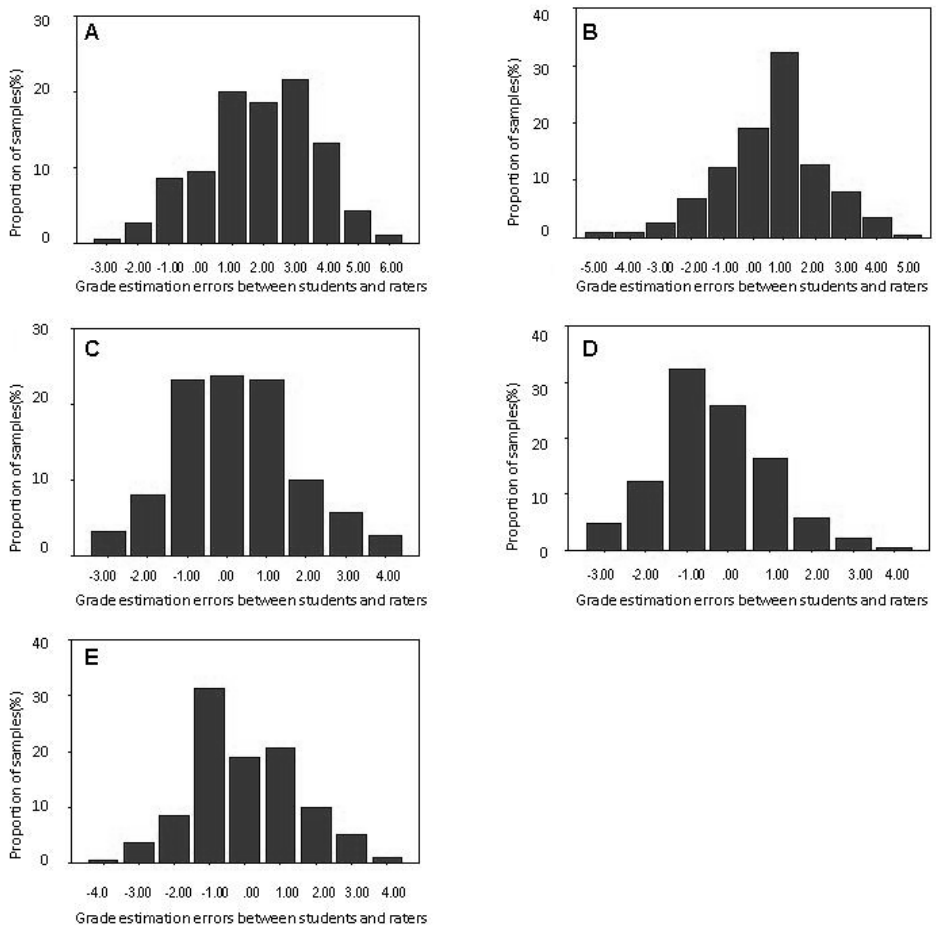

Figure 2. Analysis of Self-rating Errors. (A) Untrained Self-rating Errors by Students from Stage 1. (B) Trained Self-rating Errors by Students from Stage 2. Compared with (A), the percentage of accurate ratings after training increases radically from 10% to 19%. In addition, the percentage of overestimation (above two grades of the correct value) went down dramatically from 40.2% to 6.4 %. (C) Untrained Self-rating Errors by Students from Stage 3. (D) Trained Self-rating Errors by Students from Stage 4. (E) Grade estimation errors after two self-rating training sections from Stage 5.

The significance of change is illustrated by contrasting Figure 2(C) to Figure 2(D), which are the percentage of students' grade estimation errors for Writing Task 2.

By comparing the two graphs, though the percentage of accurate ratings (i.e. where the estimation error is zero) increases only slightly between the first and second training sessions, the percentage of overestimation (above two grades of the correct value) has dropped from 8% to 3 %. In addition, around 72% of the participating students were within one grade difference. The result from the frequencies test confirms improvement in the accuracy of the students' self-assessment for writing after the second session of rating training.

From Table 3, compared with the standard error mean of the students' first self-rating for Writing Task 1, the standard error mean of Writing 3 has decreased by 0.03. Through the paired-samples *t*-test, the result shows there is a significant difference in the students' estimation errors between the students' first self-rating of Writing Task 1 without any rating training and their self-rating of

Writing Task 3 after the two sessions of rating training, t=12.26, p< 0.01. This finding indicates that the effect of students' assessment criteria instruction on the improvement in accuracy of their self-assessment for writing is great. Among those three pairs, the most significant difference is between Writing Task 1 and Writing Task 3, which shows that self-assessment training contributes positively to the enhancement of the accuracy of students' self-assessment.

This improvement of accuracy of the students' self-assessments is made clearer by looking at the frequencies of students' estimation errors, as shown in Figure 2(E).

Compared with Figure 2(A), the percentage of accurate ratings (i.e. where the estimation error is zero) increases radically from 10% to 19% in Figure 2(E). In addition, the percentage of overestimation (above two grades of the correct value) decreases dramatically from 40.2% to 6.4 %.

All the findings point to the fact that the students' self-ratings are getting close to the ratings by the instructors after the students attended just two rating training sessions. It seems by this time they developed a significant capacity for self-assessment in CET writing. It also illustrates that self-assessment capacity is enhanced by the knowledge of assessment criteria. So the answer to the second question is that self-assessment training does increase the accuracy of the students' self-assessments.

3.3 Self-assessment Training Contributes to Language Writing

To seek the answer to the last research question, "Will self-assessment training contribute to language achievement?" Our attention is on the increases in students' writing scores over the three writing tasks.

As expected, through the paired-samples t-test, the result shows that there is a significant difference between the scores for Writing Task 1 (8.4±1.7) and Writing Task 2 (9.2±1.5), t=-7.09, p< 0.01.and that the score of Writing Task 3 (9.4±1.5) is significantly higher than that of Writing Task 1, t=-9.24, p< 0.01. In the same way, there is a significant difference between the score of Writing Task 2 and that of Writing Task 3, t=-2.180, p< 0.05. The evidence is that the students benefited from the criteria instructions, and their composition performance improved. Such findings point to the positive influence of self-assessment on the learning process, providing it is applied meaningfully and taught well.

4. Discussion and Conclusions

As a pedagogical endeavor, writing self-assessment (in terms of scores, grades, and evaluative feedback), if valid and reliable, can contribute to students' learning processes and help students enhance their writing skills. Equally important, it provides instructors with information concerning the effectiveness of their teaching, which will be seen in their students' increased writing proficiency and their achievement of the program objectives. Thus there is a good reason to get a better understanding of the nature of student writing self-assessment and of how to best assess student writing.

In the previous studies of student cognition of self-assessment (Ross *et al.*, 1999), the students reported that they paid more attention to their self-assessment because they understood the criteria. They felt ownership of the data, and they felt empowered because the instructors trusted them to rate themselves fairly.

In this study, the size of the impact of training on accuracy was large. When the instructors shared the assessment criteria with the students, the tendency to inflate grades decreased. The students received feedback on their application of the criteria which gave them a clearer understanding of College English Test Band 4 or Band 6 writing standards or classroom standards. One possible explanation might be that they had read the rating scale, so their judgments about how well they did became more accurate. Since the rating scale was available to them, they could know the criteria for their self-assessments. An alternate explanation for the strong effects of the training on accuracy was that the instructors focused on a learning objective that received extensive instructional attention in both sessions. The students became more knowledgeable about what counted in writing so that the focus on criteria contributed a lot to their understanding of what they were supposed to do.

An implication of the results of this study for writing instructors is, according to Schendel & O'Neil (1999), that "students can learn and develop the metacognitive skills they need to evaluate their own performance" reliably against given criteria, "but students do not necessarily arrive in classrooms or at college [orientated] with this ability" (p. 218), nor are any criteria they have previously used necessarily shared with a new institution.

We have come to the following conclusions. The students can perform self-assessment in writing reasonably well. Teaching self-assessment skills increase the accuracy of student self-appraisals. The self-assessment training has a positive effect on students' writing achievement.

References

Arter, J., Spandel, V., Culham, R. & Pollard, J. (1994, April). The impact of training students to be self-assessors of writing. Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).

Athanasou, J. A. (2005). Some issues with adult self-evaluations in education and training. Paper presented at the Australian Association for Research in Education, Annual Conference, University of Western Sydney, November 2005.

Hillocks, G. (1986). *Research on written composition: New directions for teachin*g. Urbana, IL: ERIC Clearinghouse on Reading and Communication Skills.

Janssen-van Dieten, A. (1989). The development of a test of Dutch as a second language: the validity of self-assessment by inexperienced subjects. *Language testing 6*/1:30-46.

Longhurst, N. & Norton, L. S. (1997). Self-assessment in coursework essays. *Studies in Educational Evaluation 23*(4), 319-330.

Oscarson, M. (1989). Self-assessment of language proficiency: rationale and applications. *Language testing 6*/1:1-13

Oscarson, M. (1997). Self-assessment of foreign and second language proficency. In C. Clapham & D. Corson (eds.), *Encyclopedia of language and education*, *Volume 7: Language testing and assessment*, 175-187. Kluwer Academic Publishers.

Ross, J. A., Rolheiser, C., & Hogaboam-Gray, A. (1999). Effects of self-evaluation training on narrative writing. *Assessing writing 6*(1), 107-132.

Schendel, E. & O'Neill P. (1999). Exploring the theories and consequences of self-assessment through ethical inquiry. *Assessing writing 6*(2), 199-227.

Shaw, S. D. & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing*. Cambridge University Press.