

# Self and Peer Evaluation of Writing in the Interactive ESL Classroom: An Exploratory Study

Dennie Rothschild  
Felicia Klingenberg

---

The evaluation of writing in the ESL classroom has traditionally been the teacher's prerogative and as such it has remained outside the interactive model of student learning. Our goal is to bring evaluation into the classroom in order to increase learners' awareness of criteria for good writing, promote greater improvement of writing by giving learners an instructional and diagnostic tool which they could use, reinforce in-process feedback with end-of-process evaluation, and foster more positive attitudes towards writing.

The students in our pilot investigation are high intermediate level adults from diverse backgrounds studying part-time during a four-month term. Our investigation is in two parts. Part one involves adapting an appropriate evaluation scale, training students in its use, and having them use the scale throughout the term to

evaluate their own and their peers' writing. In part two we study various end-of-term effects the use of the scale had on students: we test the hypothesis that students trained in the use of the scale will have a concept of good writing more congruent with that of instructors than will a control group; we compare the criteria most often cited by both groups as they judge the quality of a set of compositions; we examine the responses of both groups to a survey on their attitudes towards writing.

Our results show a slight trend in the predicted direction between the experimental group and one of the judges. We also find indications that the experimental group is using a different set of criteria in judging compositions. As well, the experimental group responds more positively to all ten statements on a writing attitude survey.

---

This paper is based on a broad investigation of self and peer evaluation of writing in the general skills ESL classroom. We define evaluation as both identification of strengths and weaknesses, as well as the actual assignment of grades based on a set of explicit criteria; we mean evaluation then in both its formative sense—defined by Cooper (1975) as “response and feedback to a writer's efforts”—and its summative sense—“finding out how much a student has grown as a writer”.

## Teacher-Centred Evaluation

Various researchers have begun to uncover problems with teacher-centred evaluation. In a study of ESL teachers' responses to student writ-

ing, typically presented in abbreviated margin and end notes, Zamel (1985) concluded, "The marks and comments are often confusing, arbitrary and inaccessible." As well, research on student responses to teacher feedback (Cohen 1987) has shown that a significant number of students misinterpret comments, or do not attend to them, even when participating in a process-oriented writing class. As Butler (1989) has aptly summed up the problem:

to put new wine into old bottles—to implement a modern approach to the teaching of written composition within a traditional system of marking and grading—can be disastrous. . . . Although many teachers have changed their approach successfully, some attempts have foundered on the hard rocks of evaluation.

### **Self and Peer Evaluation in L1 and L2**

Self and peer evaluation are well known concepts in L1 research and have become proven teaching/learning strategies in the L1 classroom. Various high school systems both in Canada (Cavanagh and Styles, 1983) and the U.S. (Katstra, 1985; Carlson and Roellich, 1983; Takacs, 1987) routinely use self and/or peer evaluation in their language arts classrooms. College teachers are employing it in freshman composition classes (Boss, 1988) and in content courses such as psychology and theatre history. Teachers of business and technical writing are finding the strategy useful (Bishop, 1987; Selfe, 1981), as are teachers of college study skills courses (King and Stahl, 1985) where students are peer-evaluated on their note-taking. So prevalent has the use of peer evaluation become in fact that a computer-assisted peer grading system called Peerate has been developed (Borchardt, 1980). Peerate is a system that allows students to rate each other anonymously while at the same time the program calculates their rating ability.

In the ESL classroom, the evaluation of writing also has traditionally been the teacher's prerogative; as such it has remained outside the current interactive model of student learning. Even after process-oriented approaches to writing instruction were introduced to ESL classrooms and students became accustomed to being given various checklists or heuristics to help them in the task of responding to each other's work, the teacher remained the only reader with grading authority. As a result, few reports by either teacher-practitioners or researchers on the use of peer and/or self evaluation as we have defined it are available. Two studies have been done on in-process feedback—comparing results between drafts—and these have reached inconclusive results on the benefits of peer feedback over teacher feedback. Partridge (in Chaudron 1983) found greater improvement with teacher feedback, but concluded nevertheless that peer evaluation may prove over time to enhance student confidence in their judgments and

sensitivity to their audience. Chaudron (1983), despite finding “no overall difference between improvement based on teacher or peer feedback”, still felt that the more the writing process can be learned “as an interaction between writers and their readers, the more the L2 learners will appreciate the functions, savour the fruits, as it were, of their newly acquired writing proficiency.”

While pursuing the benefits of peer evaluation, researchers have not failed to note that, in the ESL classroom, “students are cautious about the value of peer feedback as a source of aid in revising their writing” (Chaudron, 1983). Davies and Omberg (1987) list three disadvantages perceived by students, including lack of expertise, faulty corrections and fear of hurting each other’s feelings. Overall, however, the students in their study were more inclined to comment on the advantages than on the disadvantages.

### **Goals**

By sharing the responsibility for end-of-process evaluation (grading) with our students, we wished

1. to share with our students criteria for good writing
2. to promote greater improvement of writing by giving learners an instructional and diagnostic tool which they could use
3. to reinforce in-process feedback with end-of-process evaluation to see whether this would improve students’ attitudes, motivation and abilities to help each other throughout the drafting process
4. to foster more positive attitudes towards writing

### **Students**

The students in our investigation are high intermediate level adults from diverse language, educational and cultural backgrounds. They are studying integrated English language skills part time at a Canadian community college. The students’ goals range from general language improvement to preparation for entry to an English for Academic Purposes (also known as English for Higher Education) programme. The writing marks the students receive during and at the end of the term are often meaningless to them other than in the pass-fail sense. They are not party to how the marks were obtained, or on what they are based, and therefore don’t know where they have “gone wrong” or how to improve.

### **Method**

Our pilot investigation is in two parts. Part one involves adapting an appropriate evaluation scale, training students in its use and having them use the scale throughout the term to evaluate their own and their peers’

writing. In part two we study various end-of-term effects that the use of the scale had on students: we test the hypothesis that students trained in the use of a scale (the experimental group) will have a concept of good writing more congruent with that of instructors than will a control group; in addition, we compare the criteria most often cited by both groups as they judge the quality of a set of compositions; finally we examine the responses of both groups to a survey on their attitudes towards writing.

## Part One

The evaluation scale (see Figure 1) was adapted from the ESL Composition Profile and the Carol Sager scale to suit the requirements of our programme. Our scale is most similar to the ESL Profile in its language use and mechanics components, whereas it is most similar to the Sager scale in its content and organization categories. Sager's focus on reader awareness and elaboration of detail is more appropriate for the expressive/reflective (narrative/descriptive) kinds of writing done in our programme. Since the adapted scale was designed as part of a larger writing project focusing on content, organization and language use, we felt that a vocabulary component could be a complicating factor in student use of the scale and therefore it was not included, as it is in the ESL Profile and the Sager scale. The weighting of the categories parallels that of the ESL Composition Profile; i.e. content is weighted most heavily, followed by language use, organization and mechanics. Figure 1 also shows three categories containing 4 levels and one category with 3 levels with each level having a range of marks. As well, figure 1 shows the columns for the writer's self-assigned grades, peer-assigned (peers 1 and 2) grades, teacher-assigned grades, and a final mark, which is an average of all of the above. Each student writer kept a copy of the scale for his own records. After receiving marks from the two peers, the writer added these to his copy. The writer then gave the entire portfolio to his teacher.

Figure 1  
**Rating Scale** for composition no. \_\_\_\_\_

### **CONTENT: ideas, information or message**

- 27-35 The reader can see & feel what the writer sees & feels • ideas & details create an impression on the reader • all ideas are clear & fully developed • all ideas are nicely related to each other & to the title • details fill out the ideas & make people, places &/or events come alive • all questions are answered (a sense of completeness)
- 18-26 the reader begins to see & feel what the writer does • ideas & details start to make an impression • most ideas are clear & well-developed • most

ideas are related to each other & to the title but some details don't belong  
• details begin to fill out the ideas & make people, places &/or events come alive • a few questions are unanswered because some details are missing

- 9-17 the reader doesn't see & feel what the writer does • ideas & details make little or no impression • some ideas are clear & developed • some ideas & detail are not related to each other & to the title • there aren't enough details to fill out the ideas; as a result, people, places &/or events don't come alive • many questions are unanswered because of missing details
- 0-8 the ideas make no impression on the reader because they are confusing, hard to follow, unclear &/or undeveloped • ideas & details don't seem to be related to each other & to the title • important questions are unanswered because there are no or almost no details

### **ORGANIZATION: the arrangement of ideas in order**

- 19-25 the introduction is interesting: it makes the reader want to continue reading  
• the conclusion helps the reader understand the writer's point of view &/or feelings • a main idea ties all story parts together in an obvious logical order • each paragraph has only 1 main idea or purpose & all details support that idea
- 12-18 there is an introduction but it doesn't grab the reader's attention • there is a conclusion but it doesn't help the reader understand the writer's point of view &/or feelings • a main idea ties all story parts together but some events are told out of order • some paragraphs have more than 1 main idea or purpose, or no obvious main idea or purpose
- 6-11 the introduction/conclusion is not useful or interesting • there is a main idea but many events are out of order • many paragraphs don't have 1 main idea or purpose
- 0-5 no introduction or conclusion • the reader doesn't see point to story because ideas are so disorganized

### **STRUCTURE: the way language is used**

- 23-30 clauses are joined in the most effective/meaningful way by connectors (e.g. if, so . . . that, although, because as, since, who, which, where, whose, that, when, etc.) • each sentence is complete (no run-ons or fragments) • there are almost no errors of agreement, tense, articles, word order, word form, prepositions, etc.
- 15-22 clauses are not always joined in the most effective way but the ideas are still easy to understand • almost all sentences are complete (few run-ons or fragments) • there are some errors of agreement, tense, articles, word order, etc. but we can easily understand the story
- 7-14 clauses are not joined or are poorly joined so that ideas are sometimes difficult to understand • some sentences are complete (some run-ons or

fragments) • there are many errors of agreement, tense, articles, word order, etc. and sometimes they make the story difficult to understand

- 0-6 ideas/meaning is unclear or difficult to understand • few sentences are complete (many run-ons or fragments) • dominated by errors of agreement, tense, articles, word order, etc. which make the story confusing

**MECHANICS: the way the writing looks**

- 7-10 almost no errors of spelling, punctuation, capitalization or paragraphing • handwriting is legible
- 4-6 some errors of spelling, punctuation, capitalization or paragraphing but this does not usually interfere with understanding the story • handwriting is usually easy to read
- 0-3 many errors of spelling, etc. which may make the story difficult to understand • handwriting is hard to read

---

	WRITER	STUDENT 1	STUDENT 2	TEACHER	AVERAGE
<b>CONTENT</b>	_____	_____	_____	_____	_____
<b>ORGANIZATION</b>	_____	_____	_____	_____	_____
<b>STRUCTURE</b>	_____	_____	_____	_____	_____
<b>MECHANICS</b>	_____	_____	_____	_____	_____
<b>TOTAL</b>	_____	_____	_____	_____	_____

---

Adapted from the “ESL Composition Profile” and the “Sager Scale”

**Training the students to use the scale**

Thirty-one students in two classes began the training session by discussing the descriptors in the content category of the scale. Then each class was asked to read and evaluate a sample composition written by a student from the other class. We felt this exchange of papers between classes would be less intimidating and would allow for freer discussion. Because both classes had done the same writing task, the topic was familiar to all students. We then displayed the results on the blackboard and continued with more discussion. This procedure was repeated for the organization category. At this point we did not train the students to the language use or mechanics categories of the scale. Figure 2 shows means and standard deviations resulting from the initial training session.

Figure 2  
**Mean Scores and Standard Deviations**  
 Training session for 2 Upper Intermediate classes

	<u>content</u>	<u>organization</u>
Class 1 n = 12 composition A	x = 18.91 s = 3.29	x = 15.04 s = 3.48
Class 2 n = 15 composition B	x = 25.46 s = 3.52	x = 17.16 s = 1.83
Class 2 n = 16 composition C	x = 16.37 s = 6.35	x = 12.53 s = 3.56

Note:

- Composition A Teacher's mark = 20 (content)  
= 19 (organization)
- Composition B Teacher's mark = 23 (content)  
= 20 (organization)
- Composition C Teacher's mark = 12 (content)  
= 14 (organization)

On this first attempt, although there are some exceptions, the majority of students agreed with an acceptable range with each other and with the teacher. Because so many of the students were able to do this in their first training session, we are encouraged to believe that the use of the scale is within the students' capabilities. As a follow-up, we trained another class (a Lower Advanced class) in a similar manner. The results shown in figure 3 reveal a greater degree of conformity.

Figure 3  
**Mean Scores and Standard Deviations**  
 Training session for a Lower Advanced class

	<u>content</u>	<u>organization</u>
Class 2 composition D n = 20	x = 23.44 s = 3.38	x = 14.45 s = 1.76
Class 2 composition E n = 17	x = 21.47 s = 2.83	x = 19.25 s = 1.13

Note:

Composition D Teacher's mark = 20 (content)  
= 18 (organization)  
Composition E Teacher's mark = 26 (content)  
= 19 (organization)

At this point we are not sure if the higher agreement in the Advanced class is due to students' increased proficiency, the teacher's increased skill and confidence in the training procedure, or some other factor.

The next step was for the two high intermediate classes to use the scale throughout the term to assess their own and their peers' work. Most students had the opportunity to use it three times: that is, at the end of each writing cycle during the term. Their option was to either revise again or keep their performance in mind when beginning the next composition cycle. Some students were seen referring to the scale to peer-revise draft 1 of the following composition. When this occurred, it was in addition to the regular revision heuristic.

## **Part two**

### *Grading compositions*

Part two of the project deals with end of term results. As part of our pilot investigation, we asked students from the experimental group (E) described above and a control group (C) ( $n = 26$ ) to judge the quality of a set of 14 compositions of varying quality in order to confirm our hypothesis that students trained in the use of a scale would have a conception of good writing more congruent with that of teacher-judges. The two teacher-judges used the scale to grade the compositions. However, because the control group had never seen the scale, or indeed any scale, both the control and the experimental groups were simply asked to assign a global mark from 1 to 10 (ten being highest) without using any explicit criteria.

We used a Spearman rank order correlation to analyze the results. Most of the correlations were low (in the .30 to .37 range) and significant at or approaching the 20% level. There was a higher correlation in one case—between the experimental group and judge 2 where  $\rho = .47$ , which a two-tailed test shows to be significant at the 10% level. While the results do not confirm the hypothesis, the existence of a higher correlation in one case leads us to believe that the relationship merits further investigation.

### **Criteria cited by each group when grading**

In addition to assigning grades to the sample of 14 compositions, students in both groups were asked to write open-ended comments on the



quality of the compositions because we wanted to compare the criteria most often cited by each group. An analysis of their comments (Figure 4) shows the two groups may have used different criteria; specifically, the experimental group focused on content and organization (77% of their comments versus 34% for the control group) whereas the control group relied on structure as the basis for judging quality (33% of their comments versus 4% for the experimental group).

Figure 4  
**Criteria cited by the experimental and control groups of students  
as they judged the quality of a set of 14 compositions**

	E (n = 31)*	C (n = 26)**
Content	53% (150)	29% (78)
Organization	24% (69)	5% (13)
Combined C&O	77% (219)	34% (91)
Structure	4% (11)	33% (88)
Mechanics	6% (17)	9% (24)
Combined S&M	10% (28)	42% (112)
Other (general)	13% (38)	23% (62)

\* number of comments = 285

\*\* number of comments = 265

Note:

The *other* category contains generalized comments such as “good”, “bad”, “I like it”, “I don’t like it”.

Figure 5 gives an indication of the kinds of comments cited. We note some similarity between the experimental group’s comments and the descriptors from the scale and conclude that some of the criteria appear to have been internalized. An as yet unanswered question is, “Is there a relationship between the criteria cited by students and their ability to evaluate (their own) compositions and make improvements?”

Figure 5  
**Samples of the experimental group’s open-ended comments (edited)**

**Content**

I don’t see and feel what the writer does.

Not bad but needs more details.

Too simple description.

The ideas aren’t clear. It’s a little bit confused.

There are many things missing.  
 Should tell some more information.  
 The ideas make no impression on me.

**Organization**

Not good order.  
 The introduction and conclusion aren't strong enough.  
 Too many short paragraphs.  
 Didn't separate into paragraphs.

**Survey on attitudes toward writing**

Finally we examined the responses of both student groups to a survey on their attitudes toward writing. In the survey, students were asked to agree or disagree with ten statements. In figure 6 we see the experimental group agreeing more to all 10 statements on a writing attitude survey.

Figure 6  
**High Intermediate questionnaire on writing and grammar improvement**  
 Experimental (E) group (n = 33) and Control (C) group (n = 31)

		E %	C %		
		agree	agree	difference	rank
1	My writing is better than in September 1988.	93.94	93.55	+ 0.39	10
2	I have more ideas for writing now.	96.88	83.87	+ 13.01	3
3	Writing is easier for me.	56.67	51.72	+ 4.95	8
4	My writing is better organized.	83.87	64.52	+ 19.35	1
5	I feel more comfortable about writing now.	76.67	66.67	+ 10.00	6
6	My writing has more details now.	86.67	74.19	+ 12.48	4
7	I understand the kinds of mistakes I make now.	86.67	77.42	+ 9.25	7
8	My grammar has improved.	83.87	73.33	+ 10.54	5
9	I have more vocabulary now.	78.13	64.52	+ 13.61	2
10	I enjoy writing now.	66.67	62.07	+ 4.60	9

Note: In the few cases where a student did not respond to one of the statements, that student was not included in the tally for that statement.

A comparison of the differences between the means for the two groups for all 10 statements indicates that the two groups differed significantly in their attitudes towards writing with the experimental group showing significantly more positive attitudes (t-test = 16.76, p<.01, two-tailed). The

largest differences in agreement between the two groups concerned the areas of organization, vocabulary development, ideas, and details. It appears that the experimental group may have been influenced by the criteria from the scale in three of the four areas above: organization, ideas, and details.\*

As a follow-up, the same survey was given a term later to a Lower Advanced class trained in the same way. To the survey we added questions concerning peer revision, editing and evaluation, as well as conferencing with the teacher. Figure 7 shows that only half felt that revising with their peers helped them. Given the goals of the study, we were disappointed that the revision procedure was not perceived to be more helpful. However, a majority of students (78%) felt that evaluating other students' compositions helped them either some or a lot and many (67%) felt quite confident about evaluating other students' compositions. We were not surprised that the great majority (88%) of students perceived the conference with the teacher to be very helpful as this was in line with other studies (Davies and Omberg, and Chaudron).

\* We were puzzled by the experimental group's belief that their vocabulary had improved since we deliberately had not included that component in our scale, and had not pre-taught vocabulary. Two causes seemed most likely. One was that the students, showing interest in a major international current event, the Ben Johnson affair, helped the teachers select the topic. The other was a jigsaw reading activity in which, as preparation for the writing task, the students taught each other the content and vocabulary related to the above topic.

Figure 7

**Lower Advanced questionnaire on writing and grammar improvement**  
(n = 18)

		agree	disagree
1	My writing is better than in January.	17	—
2	I have more ideas for writing now.	17	—
3	Writing is easier for me.	12	6
4	My writing is better organized.	15	2
5	I feel more comfortable about writing now.	15	2
6	My writing has more details now.	16	1
7	I understand the kinds of mistakes I make now.	15	2
8	My grammar has improved.	13	4
9	I have more vocabulary now.	13	2
10	I enjoy writing now.	12	2

### Which of the following activities helped you?

	a little	some	a lot
1 my oral revision group	8	4	5
2 my editing group	4	9	3
3 conferencing with my teacher	—	1	15
4 evaluating other students compositions	3	9	5
5 in-class grammar exercises from stories	2	6	9
6 corrections from students compositions (on the overhead projector)	3	10	3

### How confident do you feel about evaluating other students' compositions?

2	8	4
---	---	---

We intend to continue administering this survey to other High Intermediate and Advanced classes in the future to see if these positive responses to peer evaluation continue.

### Conclusion

In conclusion, we find a slight trend in the predicted direction between the experimental group and one of the judges; we find indications that the experimental group is using a different set of criteria in judging compositions than the control group; in addition, we see the experimental group giving more positive responses to all ten statements on a writing attitude survey. Because results are encouraging but not conclusive, we feel that further research should be undertaken.

However, there are problems inherent in this kind of classroom research that complicate the search for conclusive answers. The fact that the students who participate in rater training sessions are also the students whose work is to be judged means they are not completely objective. To a greater or lesser degree, they may be comparing their work to the work to be rated in the training sessions, and this may affect how they view the training compositions. Also, especially at the beginning of a term, not all students see the value in participating in the process expected of them. This resistance may affect how well they perform within it.

Many questions remain to be answered. How does the use of the scale as a diagnostic/instructional tool inform students' writing development? How does the use of the scale affect the revised product? In other words, how does the use of the scale to evaluate their own and their peers' writing help students 1) see problems in their own writing and 2) attempt to solve these problems. As well as investigating the impact the use of the scale has on learning and learners' strategies, we would be well advised to study its impact on motivation and attitude towards writing. Finally, and perhaps

most germane to students becoming independent writers, can (and if so, how can) ESL students be guided to establish their own criteria for the kinds of writing necessary in their lives?

Further qualitative and quantitative research will provide a better understanding of the merits of self and peer evaluation as a tool to improve writing in the ESL classroom.

---

#### NOTE

We'd like to thank Dr. D. Allison and Dr. A. Cumming of the University of British Columbia for statistical help, advice, comments and questions. Any remaining problems are the responsibility of the authors alone.

#### REFERENCES

- Bishop, Wendy. (1987). Revising the Technical Writing Class: Peer Critiques, Self-Evaluation and Portfolio Grading. Paper, Annual Meeting of the Pennsylvania State Conference on Rhetoric and Composition, State College, PA. ED 289 178.
- Borchardt, Donald. (1980). A Computer Assisted System for Grading Papers. ED 246 851.
- Boss, Roberta. (1988). Formative Evaluation of College Composition: A Formula for Revision and Grading. ED 289 554.
- Butler, Sydney. (1985). New Bottles For New Wine: Evaluation in a Modern Writing Program. *English Quarterly* 18.
- Cavanagh, Gary and Ken Styles. (1983). Evaluating Written Work. *English Quarterly* 16.
- Carlson, Diana and Carol Roellich. (1983). Teaching Writing Easily and Effectively to Get Results. Part II: The Evaluation Process. Paper, Annual Meeting NCTE Spring Conference. Seattle, Washington.
- Chaudron, Craig. 1983. Evaluating Writing. Effects of Feedback on Revision. Revised version of a paper presented at the Seventeenth Annual TESOL Convention. Toronto, Canada.
- Cohen, A. D. (1987). Student Processing of Feedback on Their Compositions. In *Learner Strategies in Language Learning* (Ed: A. Wenden, J. Ruben). Prentice-Hall.
- Cooper, Charles. (1975). Measuring Growth in Writing. *English Journal* 64.
- Creighton, James L. (Ed). (1987). A Potpourri of Practical Ideas. ED 281 207.
- Jacobs, Holly, et al. (1981). *Testing ESL Composition: A Practical Approach*. Rowley, Mass.: Newbury.
- Katstra, Joyce. (1985). The Effects of Peer Evaluation on Attitude Toward Writing and Writing Fluency of Ninth Grade Students. ED 268 581.

- King, James and Norman Stahl. (1985). Training and Evaluating Notetaking. College Reading and Learning Assistance Technical Report, 85-06. ED 263 537.
- Sager, Carol. (1973). Sager Writing Scale. From Ed.D Dissertation, Boston University. ED 091 723.
- Selfe, Cynthia. (1981). Using Groups to Pre-Evaluate Papers in the Technical Writing Classroom. ED 226 369.
- Takacs, Claudia. (1987). AWE: Classroom Use of a State Testing Program. *English Journal* 76.
- Zamel, Vivian. (1985). Responding to Student Writing. *TESOL Quarterly*, Vol. 19, no. 1.

### **THE AUTHORS**

Dennie Rothschild, co-author of *Oral and Written Composing* published by Vancouver Community College, is an instructor in the ESL and TESL (Writing) programme at the King Edward Campus of VCC. She is also completing a Master's thesis at UBC. A former co-ordinator in the English Language Skills Department at VCC, she is current President of B.C. TEAL.

Felicia Klingenberg, ESL instructor, has an MA in English from the University of Waterloo and has worked as a journalist. She is studying linguistics at UBC as well as teaching in the English Language Skills Department at VCC.