Advances in Language and Literary Studies
Vol. 3 No. 1; January 2012

Grade Expectations: More than Meets the Eye

Arshad Abd Samad

(Universiti Putra Malaysia, Malaysia)

Zamzam bt Ahmad

(Universiti Malaysia Pahang, Malaysia)

doi:10.7575/aiac.alls.v.3n.1p.69

Abstract

Raimes (1983) has identified nine components necessary to produce a piece of writing that is clear, fluent and effective. These are also the aspects that are considered when assessing writing. The common practice is to have raters score the essays and they are provided with a rating scale for this purpose. A training and practice session is also included. A consensus is usually the objective but McNamara (1996) comments that training has a limited effect on raters. This paper is an attempt to find out how teachers perceive "good writing" and how their perception influences the outcome of the rating procedure. To do so, ten English Language teachers from various backgrounds are asked to rate a short essay of about 150 words. They are also asked to complete a questionnaire on their beliefs about writing assessment and this is followed by an interview to elicit a more extended response on the area.

Introduction

The direct assessment of writing is a widely used method of assessing students' writing in Malaysian public examinations from the primary to the pre-tertiary level. Various parties and stages are involved from developing the writing tasks to finalising the grades. Shaw and Weir (2007) in relating the grading process adopted by ESOL Cambridge identify these stages as (i) task development, (ii) test administration, (iii) scoring and (iv) grading and awarding. It can be assumed that a similar procedure takes place here in Malaysia as well.

Usually each group of personnel involved in the assessment process works in isolation meaning there is little or no interaction among the four groups. This is even more so with those involved in the lowest rung of the scoring process – the teacher-rater. This group is

tasked to rate the essays based on a criterion that has already been agreed on by the examination governing body concerned and its top personnel.

A team or group leader handles the meetings, leading a team of between six to ten members. Meetings or a "grade coordination" exercise is conducted to ensure an acceptable degree of consensus and consistency. The raters are instructed to use either benchmark scripts or there are descriptors in the rating scale that they can use as "hooks" or indicators. Both ways do not always help because benchmark scripts do not always fall neatly into one band or category. In addition, actual scripts tend not to explicitly display the "hooks" or indicators.

It has been suggested that raters sometimes have their own personal and different interpretations that they may fall back on whenever they encounter "problematic" scripts. As such, it can be said that how the raters rate may not be a straightforward matter of assessing what they read but also how they "interact" (Cumming, 1990; Lumley, 2005) or "engage" (Charney, 1984) with and "react" to the text. As argued by Lumley (2005, p.27), the rater not only interprets the text but also interacts with the text and "... other various features of the situation notably the task and text type and the task topic, all of which contribute to the quality of writing and perceptions of this quality."

Since consensus among the raters is important for inter-rater reliability, perhaps it should be determined what makes raters similar. On the other hand, if they are different, then it would be worthwhile to find out the reason so that this issue can be addressed during rater training.

ESL writing ability: The prescribed traits

Assessing writing involves looking at how well a candidate has presented the traits involved in the writing construct and how the teacher/rater assesses the candidate's response. Before the teachers' perception of writing is considered, it may be useful to see how their views as compared with the experts' and to see how these traits are expressed as descriptors in the rating scales.

Raimes (1983, p.6) identifies nine components in ESL writing which are (i) content, (ii) the writer's process, (iii) audience, (iv) purpose, (v) word choice, (vi) organisation, (vii) mechanics, (viii) grammar and (ix) syntax. All these elements work in conjunction to

produce writing that communicates ideas that are clear, fluent and effective. Murray (1982) and Scott (1996) have added another element which is the writer's voice or perspective.

These traits are usually described in terms of accuracy and fluency and used in either holistic or analytical rating scales. They may be given equal or unequal weightage depending on the purpose of the test. Although the traits that can be assessed are extensive, Cohen (1994) maintains that only a selection is included at any one time. This is possibly due to the time allocated for assessment, relevance of the traits to the task set, the level of difficulty to assess a particular trait and cost.

Teachers' rating style

Generally, teachers have the knowledge about the elements that constitute quality writing but this knowledge may not always transfer to what they actually practise. Some may place a higher importance on certain features and would attend to or focus on these while rating. Research has indicated that personal beliefs may vary and they can influence the teachers in their decision making process while rating essays. This is regardless whether they are working with a specific scale or whether training has been given.

In a study conducted by McNamara (1990), grammar appears to be the trait that the raters focussed on although the objective of the test – the Occupational English Test or OET – was to measure communicative ability. He reported that "... candidates were to some extent measured ... on selected features which were important to the raters, independently of the design of the test" (McNamara 1990, p.69)

A study carried out by Cumming (1990) involved experienced and inexperienced ESL teachers rating without a scale. The two groups were found to be most similar in terms of language use which considered syntax and errors. This may be accounted for by language rules that are already set and are not open to interpretation. Cumming also added that as a group, 30% of the behaviour of the inexperienced group involved error correction. In contrast, the experienced group preferred to classify the errors.

In a "follow up" study but involving only experienced raters by Cumming, Kantor and Powers, (2002), they reported that the raters stated they were influenced by their previous experiences as raters or ESL instructors. These influences however, varied from one rater to

the next. Cumming and his co-researchers argued that these raters may have found it difficult to "unlearn" their rating behaviour that they have become skilled at using. If this behaviour can actually be "learned", then there may be fewer problems about re-training the teachers. Furthermore, it could also be argued that the more rating experience the teachers have, the more adept they become.

In a more recent study, Barkaoui (2007) found that the four raters involved adopted different rating processes regardless of the rating scale (holistic or analytical) they were using. The think aloud protocols revealed that the raters attended to different traits while rating. The researcher added that despite being instructed to use the scales, all four raters unavoidably referred to what they teach and to their expectations of the candidates' responses while rating.

The studies mentioned tend to involve ESL teachers who are native speakers of English, or reported to have a high proficiency of English and/or are experienced ESL teachers. To find out whether the Malaysian ESL teacher-rater would have a personal rating style, a study was carried out. The study and the preliminary findings are presented next.

Teachers' perception of writing ability

Ten teachers were involved in this on their opinions and beliefs about writing ability and how these opinions and beliefs would influence their rating style. The teachers were selected for their variability to examine whether despite the differences in their opinions, beliefs and backgrounds they would report any similarities. To gather information about the teachers, they were asked to complete a questionnaire. This was followed by a rating session and an interview.

The questionnaire consisted of three sections:

a. *personal information* where they were asked about their academic qualifications
b. *teaching experience* where they were asked about the level(s) and the duration they have been teaching as well as the subject(s) they have taught and are teaching. They were also asked to share their experience with learning and teaching writing.
c. *beliefs about writing* where they were asked to rank the descriptors commonly associated with writing as very important (VI), important (I), less important(LI),

not important (NI) or do not know or unsure (O). In addition, they were also asked about their rating experience and their knowledge of rating scales.

A rating session was conducted where the teachers had to rate a short introductory paragraph on *"Surrogacy"* written by three different students. The teachers did not work with a rating scale but were instructed to rate as they would normally do with their own students. The intention was to find out the teachers' personal opinion and to see how it matches the descriptors they have earlier ranked in the questionnaire.

The rating session was followed by a short interview where the teachers were asked about: (i) what constitutes good/quality writing, (ii) their comments about the essays they have rated and (iii) information given in the questionnaire which needed clarification. It is felt that clarification was required to assess the opinions the teachers voiced were actually the same but expressed in different ways (Cummings et al., 2002). It is felt that the interview would be able to reveal a more complete picture about the teachers' opinion and help prevent a wrong inference (about what they had written) from being made. The information elicited through the questionnaire, rating session and interview is presented below.

Writing ability and rating: what ten teachers say

The questionnaire

The ten teachers (identified as A to J) have different levels of qualifications from a first to a masters degree. They also came from slightly different fields of studies – linguistics, English and ESL. They have teaching experience ranging from more than two years to over thirty years. Some have taught at different levels starting from the primary to the tertiary levels. Six have had rating experience for at least one of the Malaysian public examinations.

None of the descriptors listed were judged not important and none of the teachers expressed uncertainty in their perception of the descriptors. The information from the questionnaire indicated that all the teachers agreed there must be clear, fluent, and effective communication of ideas. However, they differed slightly in how to achieve this aim as the teachers evaluated the rest of the descriptors with a slightly different order of importance.

Table 1: Teachers' ranking of writing descriptors

Item	Descriptors	VI	I	LI
a.	Clear, fluent, and effective communication of ideas	***10***		
b.	Has a clearly expressed main idea/thesis	***7***	3	
c.	The purpose is clearly shown	***6***	4	
d.	Clear and relevant content	***6***	4	
e.	Ideas are cohesively expressed	4	6	
f.	Ideas are well supported with details/elaborations and/or examples	5	5	
g.	Original/creative ideas	4	5	1
h.	Sentence structures shows variety in length and structure	1	8	1
i.	Sentence structures are accurately used	3	7	
j.	Sentence structures are appropriately used	5	4	1
k.	Sentence structures are effectively used	2	6	2
l.	Accurate grammar	***6***	4	
m.	Accurate spelling	4	5	1
n.	Appropriate register	1	9	
o.	Wide use of vocabulary	1	9	
p.	Paragraphing is effectively used/shows unity	2	8	
q.	Appropriate language use for purpose and audience	5	4	1
r.	Shows awareness of the reader's presence	3	3	*4*
s.	Correct punctuation	2	5	*3*
t.	Has appropriate tone/mood/attitude	1	9	

Looking at the information given regarding teachers' perception on which descriptors are more important than others, it appears that generally, content is perceived as being more important compared to language. This is because those descriptors related to language such as variety, as well as appropriate and effective use of sentence structures were deemed important rather than very important. This was despite more descriptors related to language were listed compared to content.

The descriptors commonly perceived as being *very important* were: (i) a clearly expressed thesis and main idea (seven teachers), (ii) purpose is clearly shown (six teachers), (iii) clear and relevant content (six teachers) and (iv) correct grammar (six teachers). Those that were regarded as *less important* were: (i) awareness of the reader's presence (four teachers) and (ii) punctuation (three teachers). Very few of the descriptors fell into this second category.

For four teachers – A, B, D and F, none of the descriptors were deemed as being less important. Concerning the most important aspect when they rate an essay (question 13), the common threads are "idea" and "maturity". Some of the comments made are *maturity of thought* (Teacher A), *maturity in writing* (Teacher B), *maturity of ideas* (Teacher F), and *clear, cohesive flow of ideas* (Teacher I). Teacher A explained her stand by saying that being a sixth form teacher, she expected her students to be well read to keep up with current issues.

Only Teacher E stated grammar teachers' responses were consistent with the way the teachers ranked the descriptors listed.

For the ten teachers, language elements were not viewed as important as content. The language aspect that was seen as very important was correct grammar, followed by the appropriate use of sentence structures and appropriate use of language for purpose and audience. Other language elements like register, vocabulary, tone and punctuation were generally ranked as important.

Apart from ranking the items, the teachers were asked about how they dealt with "difficult" scripts. Studies that have been cited previously have suggested that teachers tend to turn to their own personal beliefs to what is most important to them when they encounter such scripts. As nine of these teachers had stated "ideas" as the most important element, so it was expected that a similar response would be articulated here. The seven teachers who answered this question did show a tendency to look for the content or ideas in this type of scripts.

Teacher B stated that she would try to "*find any idea that can be accepted*" and for Teacher D, she would choose "*the one with better justification ... and maturity of thought*. Holistic or impression marking was the choice stated by Teacher G and Teacher I. For Teacher E and Teacher F, scripts which were accurately written but lack the required words (hence, lacking content) or interest was a problem. Both felt such scripts should be penalised and be given just an average score. Teacher A said she had had no problems as the rating scale she usually used could adequately address such issues.

Apart from this, the teachers were asked about their understanding of the different types of scales and whether they had a preference for a particular scale. They appeared to have a general idea about the scales. Four teachers – A, F, G and H indicated they had no preference because they felt that the type of scale used should be appropriate to the task. Five teachers – B, C E, I, and J preferred the analytical scale citing it to be "*more accurate", "focussed*", "*easier to rate", "easier to locate scores"* , while the most experienced teacher-rater in this group (Teacher E) made a diagnostic comment – "*can evaluate weakness."* As for Teacher D, the holistic scale is preferred.

It can be said that both Teacher D and Teacher E were actually thinking about the purpose of a test when choosing a scale, making them similar to the four teachers mentioned earlier. If this is so, then they showed a pattern which made them similar to the group with the same rating style described in the next section.

Rating session

If the teachers were categorised according to their rating experience, there would be three groups which are: (i) highly experienced raters, (ii) the less experienced raters and (iii) those who have never rated public exams before. However, if the teachers were grouped according to their rating styles, there seems to be five "patterns" where the teachers' comment showed a tendency for different foci – on language, content or both. Two did not make any comments.

The five groups are as follows:

- identified and made comments about content, language and organisation (Teachers A,D,E,F, G and H)

  For example:

  Purpose of writing not clear (Content, Teacher A, overall)

  Flimsy structures (Language, Teacher E, in text)
- identified and made comments about language errors (Teacher J)

  For example:

  Use of simple and compound sentences (Language, Teacher J, overall)

  Errors in sentence structures (Language, Teacher J, overall)
- identified language errors but made comments about content (Teacher C)

  For example:

  No elaboration, lacks details (Content, Teacher C, overall)

  Familiarity with content (Content, Teacher C, overall)

- identified language errors only and made no comments (Teacher B)
- identified and corrected language errors but made no comments (Teacher I)

Two types of comments can be distinguished: (i) overall comments or a summary that the teachers write at the end of the essays and (ii) in text where the comments were written to address specific points in the essays.

Table 2: Types of comments

Teacher	Type	
	Overall/Summary	In text
A	√	√
B	-	-
C	√	-
D	√	√
E	√	√
F	√	√ (minimal)
G	-	√
H	√	-
I	-	-
J	√	-

Interview data

The teachers tend to respond in a similar manner when it comes to what they perceived as a good piece of writing. Comments made were related to how the ideas were presented, maturity of thought, and whether the essay had a clear voice. Language elements like grammar and structure were also seen to be important, but perhaps less important compared to content.

Five teachers – B, E, F, G and H connected quality writing with ideas and content. Teacher B said that quality writing is one that has "*good, interesting ideas"*. Teacher E and Teacher G mentioned "*ideas that are effectively communicated*" and "*clear communication of ideas*" respectively while Teacher H said that "*good writing delivers* (a) *message to* (the) *audience, Has* (a) *purpose."*

Four teachers emphasised both content and language where Teacher D felt that the ideas put forward are of prime importance – "*good, interesting ideas"* and then adding "... *must have content, language and language expression*." Where Teacher F is concerned, a good essay is "*well structured, mature,* (with the) *message clearly conveyed.*" For Teacher I, to be rated as good writing, the "*Meaning must come through. But grammar* (is) *important* (too). (Has) *good ideas.* (Must be) *Coherent and cohesive.*" Teacher J equated good writing to "*A masterpiece. Correct or almost correct grammar. Has flow of ideas*."

Teacher A made a different comment to the others when she said that a good essay would "have *a voice"* where the writer interacts with the reader. According to her, a good essay "*Speaks to the readers. Even if the readers do not agree, they will respond or create a*

response". A similar comment was also conveyed by Teacher H when she said that, "*audience awareness* (is) *important*."

Teacher C comment is more rubric oriented when she commented that it is important the student answer the question, adding that "*... good language and content useless if the student doesn't anwer the question."*She also felt that better grammar means better writing, an opinion shared by Teacher H.

Besides that four teachers commented on the rating exercise and/or the scripts. Teacher A, the most experienced rater-teacher involved, said that she could not work without a scale as having one would help her to focus. She added that, "*I am not sure about how to rate the essays. So, I have given every script a mark*." She thought the essays were more or less the same and gave them the same mark.

The other senior teacher, Teacher F, asked whether the scripts were written collaboratively because they all had similar content. This similarity was also commented on by Teacher A. Both mentioned that the essays are incomplete, stating that the essays ended by mentioning the factors the students would like to discuss but did not do so. Incidentally, they were the only two who mentioned this.

The two least experienced teachers – Teacher I and Teacher J said the essays were "c*onfusing"* and *"difficult to read*". During the rating session, Teacher I asked clarification questions about the essays while Teacher J said she needed a dictionary to help her.
The teachers were also asked to clarify the comments which they had given in the questionnaire which sounded vague or incomplete. Only four of the teachers were involved and most of the issues concerned with problematic scripts and the rating scales.

On how to deal with problematic scripts, Teacher C who said that she would consider how relevant the script was to the task set. She earlier stated her preference for the analytic scale because it was easy to use but added that it was also useful as a diagnostic tool. Teacher H said that she rated problematic scripts based on the type (multiple word errors or single word) and frequency of errors, a similar notion asserted by Teacher J who said she would look at the grammatical errors. Teacher I who stated she preferred the analytic scale commented that she

usually worked with a combined scale on a single task. The holistic scale is for language and the analytic scale is for content.

Rating writing: Beyond the text

Perception about writing ability: The descriptors

The information gathered in this study suggests that the teachers did have a preference for certain descriptors. As they themselves are ESL speakers, this may be the reason they placed more emphasis on content or ideas. It may be assumed that the teachers felt that as long as the "message gets through", then it was not as important to have "flawless" language. It may also be a reflection of their own students who were generally of average ability.

Concerning the less important aspect, perhaps this is because the teachers did not feel that punctuation was a problem for their students. It was something that their students would be able to manage on their own. Regarding reader's presence, the teachers were more or less equally divided about its order of importance. This may be because the teachers feel that their students already have too much to focus on concerning content and language that audience could take a back seat. Furthermore, the essays would have typically only one audience – the teacher. Therefore, as long as the teachers can understand their students, it is sufficient.

However, more teachers perceived the language used should be appropriate for the purpose and audience. One would think that these two items (audience awareness and appropriate language would produce the same rank. Four teachers (C,D,G and J) have ranked them differently. It may be due to how they interpreted the descriptors. As the teachers were not asked to explain how they perceived the descriptors listed, this is at best an assumption.

Perception about writing ability: The teachers

The highly experienced group of teacher-rater more or less agreed on the descriptors but showed a wider discrepancy concerning appropriateness of structures, creativity and audience. Two of the teachers (Teachers A and F) showed better consensus than the third (Teacher E). This may be because despite their similar rating experience, the third has taught for fewer years (16 compared to their 30 years).

Turning to the less experienced group, one teacher (Teacher B) stood out showing disagreement with the other two in the group. There did not appear to be a uniting factor for

all three. In fact, two of the teachers (Teachers G and H) perceived the descriptors more similarly to Teacher E in the highly experienced group. For Teachers E and G, this may be due to the similar number of years they have been teaching (16 and 19 years respectively).

The third group of teachers who were banded together because they had no previous rating experience (of standardised public examinations) also had one member (Teacher I) who had very different perceptions. There were seven descriptors for which this teacher did not share the same opinion as the others in the group. Looking at the ranking given to the descriptors, there did not appear to be a clear pattern. If in the earlier two groups teaching experience appeared to be the cohesive factor, this was not so with the third group. This is because one teacher (Teacher D) in the third group has far more teaching experience.

Where the interview was concerned, what can be noted is that teachers who have been or are raters, tend to respond with scale specific answers or by using "jargon" that is associated with such public examinations.

Rating style

All the teachers who are or have been raters belonged to one group with Teacher B being the only exception. The other new rater was Teacher H who may have made some comments because of the rating condition. It was conducted similar to a coordination meeting where the raters would do the initial rating under the supervision of a team leader. This may have compelled this teacher to add comments as this is required in such conditions.

As only the introductory paragraphs were used , Teacher A did not actually think that she would find a voice, but instead looked for something similar in commenting about the purpose being absent or unclear. Had the writing been complete, this teacher might have commented on what she perceived as highly important.

The rating done by Teachers C and J appeared to concur with what they said in the interview – that is the emphasis on grammar. On the other hand, their lack in rating experience may have required them to be informed that they should make comments on both language and content. However, as the interest is to see what teachers actually did when they rate without a scale, this instruction was not given.

Teacher C's lack of comments concerning language may be because she has identified the language errors, and felt no further explanation was required. It appeared that with this teacher, the two elements were assessed separately.

Finally, Teacher I made no comment but made extensive corrections to the students' grammar. When asked in the interview for the reason she corrected the errors, she responded "*I don't know why. It is easier for me this way.*" One would feel that her rating behaviour was that of a typical teacher who identifies all the errors in the hope that the students can make diagnostic use of the corrections later. Since she was instructed to rate the essays as she normally did, this was what she had done. Her rating style was also similar to the inexperienced ESL teachers involved in the Cumming (1990) study. Although they knew the essays would not be returned to the students, they still made extensive corrections.

Although there were similarities among the teachers, it was not surprising to find that there was no clear demarcation of the teachers. This means that the raters could not actually be differentiated easily based on the characteristics considered. This was also one of the findings reported by Cumming (1990). He found that the strategies used by the experienced and inexperienced teachers would sometimes overlap.

Types of comments

There appeared to be more comments made regarding language although the responses revealed the teachers' preference for content. However, this may be due to the scripts they had to rate which had numerous language errors that could have influenced the teachers. The limited number of ideas as it was a group task (at the pre writing stage) may have also prevented the teachers from making more comments concerning content. McNamara (1996) suggests that different tasks and different levels of proficiency may influence what the raters focus on and that could possibly be the case here.

As mentioned earlier, the teachers may also have different interpretations of the descriptors. Since the teachers were not asked to clarify which descriptors they understood as being related to content, language and organisational traits, the above is merely an assumption. Perhaps they could have been interviewed about this, so a common interpretation can be determined. As argued by Lumley (2005), differences in coding may induce different interpretations which may actually be the same.

As a scale was absent in this study, these teachers had no point of reference. The more experienced teachers and raters could depend on their prior experience. However, the least experienced ones may only be confident to make general comments. Nevertheless, they could have used the descriptors listed in the questionnaire to make comments.

Another matter to add here is that in the questionnaire both Teachers B and I have identified the most number of descriptors as being very important – 13 and 16 respectively. Judging too many descriptors as very important (or less important) may make the raters' task more difficult especially with problematic scripts. This is because the raters would not be able to find that one descriptor which can help them focus before they award a score as suggested by Eckes (2008).

Conclusion

Considering the characteristics of teachers involved, profiling them does not appear to be a straightforward exercise. It may be possible to have a group profile but a much larger and selective sample is obviously needed.

Perhaps one can start with profiling the raters according to their teaching experience and see how similar they are. Educational background can be considered but as English teachers tend to come from a wider field nowadays, it may not be as easy to find the connection. Another variable of interest is rating proficiency as considered by Wolfe (1997) and Wolfe, Kao and Ranney (1998) where their findings indicated that the more proficient raters were better at using the descriptors presented in the scale.

If there are specific characteristics that are associated with different groups of raters, it may be that this matter can be addressed during rater training. Perhaps by having a group of raters who are more uniformed, rater training can be made more effective. On the other hand as noted by Weigle (1994; 1998), although through training, raters can be taught to use the scoring guide accordingly, there are factors involved that cannot be accounted for (Lumley, 2005).

Perhaps, in order to bridge the gap between teaching and evaluation, the teacher-rater could become more involved in test development. Weigle (2002) and Barkaoui (2007) felt that

raters should be involved in scale development to ensure that the most suitable scale is produced and used appropriately and consistently.

As this study only looked at the teachers' perception of writing ability, the next step would be to examine the (cognitive) processes the teachers go through while rating. One of the ways this could be done is to use think aloud. It is hoped that through this method a clearer picture of the rating process can be obtained. This may help to ascertain the extent the teachers are similar or different and can be used to build a framework for rater profiling.

References

Barkaoui, K. (2007). Rating scale impact on EFL essay marking:A mixed method study. *Assessing Writing , 12* (2), 86-107.

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

Cohen, A. D. (1994). *Assessing Second Language Ability in the Classroom* (2nd ed.). Boston, MA: Heinle & Heinle Publisher.

Cumming, A. (1990). Expertise in evaluating second language composition. *Language Testing , 7* (1), 31-51.

Cumming, A; Kantor, R & Powers, D.E (2002). Decision Making while Rating ESL/EFL Writing Tasks : A Descriptive Framework. *The Modern Language Journal*, 86 (1), 67-96

Eckes, T (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, 25(2), 155-185.

Lumley, T. (2002). Assessment criteria in a large scale writing test:what do they really mean to raters? *Language Testing , 19* (3), 248-276.

Lumley, T. (2005). *Assessing Second Language Writing:The Rater's Perspective.* Frankfurt: Peter Lang.

McNamara, T. F. (1990). Item response theory and the validation of an ESP test for health professionals. *Language Testing , 7* (1), 52-75.

McNamara, T.F. (1996). *Measuring Second Language Performance*. New York: Addison Wesley Longman

Murray, D.M. (1982) Learning by teaching. Montclair, New Jersey. Cook Publishers

Raimes, A. (1983). Techniques in teaching writing. New York. Oxford University Press

Scott, V.M. (1996). Rethinking foreign language writing. Boston,MA. Heinle & Heinle Publishers

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

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

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing , 15* (2), 263-287.

Wolfe, E. W. (1997). Reading style and scoring proficiency. *Assessing Writing , 4* (1), 83-106.

Wolfe, E.W, Kao, C.W, & Ranney, M (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Communication.* 15(4). 465-492