Writing evaluation: what can analytic versus holistic essay scoring tell us?

Nahla Bacha*

Lebanese American University, PO Box 36, Byblos, Lebanon

Received 14 November 2000; received in revised form 26 February 2001; accepted 29 March 2001

Abstract

Two important issues in essay evaluation are choice of an appropriate rating scale and setting up criteria based on the purpose of the evaluation. Research has shown that reliable and valid information gained from both analytic and holistic scoring instruments can tell teachers much about their students’ proficiency levels. However, it is claimed that the purpose of the essay task, whether for diagnosis, development or promotion, is significant in deciding which scale is chosen. Revisiting the value of these scales is necessary for teachers to continue to be aware of their relevance. This article reports a study carried out on a sample of final exam essays written by L1 Arabic non-native students of English attending the Freshman English I course in the English as a Foreign Language (EFL) program at the Lebanese American University. Specifically, it aims to find out what analytic and holistic scoring using one evaluation instrument, the English as a Second Language (ESL) Composition Profile (Jacobs et al., 1981. Testing ESL Composition: A Practical Approach. Newbury House, Rowley, MA), can tell teachers about their students’ essay proficiency on which to base promotional decisions. Findings indicate that the EFL program would benefit from more analytic measures. © 2001 Published by Elsevier Science Ltd.

Keywords: English as a Foreign/Second Language; Writing evaluation; Essay evaluation; Analytic evaluation; Holistic evaluation

* Tel./fax: +961-9-547256.
E-mail address: nbacha@byblos.lau.edu.lb (N. Bacha).
1. Introduction

In adopting scoring instruments with clearly identifiable criteria for evaluating English as a Foreign Language/English as a Second Language (EFL/ESL) essays, the paramount guiding principle is obviously the purpose of the essay. Evaluating essays in EFL/ESL programs has been mainly for diagnostic, developmental or promotional purposes (Weir, 1983, 1990). Thus, in order for these programs to obtain valid results upon which to base decisions, the choice of evaluation instrument to be adopted becomes significant. Although Gamaroff (2000) states that language testing is not in an “abyss of ignorance” (Alderson, 1983 cited in Gamaroff, 2000), the choice of the ‘right’ essay writing evaluation criteria in many EFL/ESL programs remains problematic as often those chosen are inappropriate for the purpose. This is most crucial when decisions concerning student promotion at the end of the semester to the next English course have to be made mainly based on essay writing scores. It is then important that teachers are aware of the potential of the evaluation criteria being adopted. This article focuses on examining what two types of scoring instruments, analytic and holistic, can tell us in essay writing evaluation for promotional purposes.

2. Literature review

In evaluating essay writing either analytically or holistically, teachers have had to address a number of concerns that have been found in the research to affect the assigning of a final score to a writing product. Some of these concerns have included the need to attain valid and reliable scores, set relevant tasks, give sufficient writing time, set clear essay prompts, and choose appropriate rhetorical modes (Braddock et al., 1963; Cooper and Odell, 1977; Crowhurst and Piche, 1979; Hoetker, 1982; Quellmalz et al., 1982; Raymond, 1982; Brossell, 1983; Smith et al., 1985; Carlman, 1986; Hillocks, 1986; Kegley, 1986; Caudery, 1990; Hayward, 1990; Read, 1990; Purves and Degenhart, 1992; Tedick and Mathison, 1995; Elbow, 1999). These issues are also of concern and often require different consideration when evaluating non-native student essays in English (Oller and Perkins, 1980; Jacobs et al., 1981; Hamp-Lyons, 1986a, b, 1990, 1991, 1995; Horowitz, 1986a, b; Lloyd-Jones, 1987; Kroll, 1990a, b; Tedick, 1990; Pere-Woodley, 1991; Reid, 1993; Cushing-Weigle, 1994; Douglas, 1995; Shohamy, 1995; Upshur and Turner, 1995; Kunnan, 1998). However, two main related concerns in both L1 and L2 essay evaluation literature are the appropriateness of the scoring criteria and the standard required (Gamaroff, 2000). Kroll (1990a) emphasizes the complexity in setting criteria and standards by which to evaluate student writing.

---

1 Evaluation sometimes refers to assigning a score to a direct writing product based on predefined criteria. It is distinguished from assessment in that the scoring for the latter focuses more on feedback and alternative evaluative techniques in the process of learning. In this study, the term evaluation will refer to assigning a score to a direct writing product.
There is no single written standard that can be said to represent the ‘ideal’ written product in English. Therefore, we cannot easily establish procedures for evaluating ESL writing in terms of adherence to some model of native-speaker writing. Even narrowing the discussion to a focus on academic writing is fraught with complexity. (p. 141)

This difficulty, however, is best addressed by identifying the purpose of the essay writing in EFL programs prior to adopting any criteria. Once the purpose is determined, two important related issues to be considered are the content and construct validity of the essay task.

A writing evaluation instrument is said to have content validity when “...it evaluates writers' performance on the kind of writing tasks they are normally required to do in the classroom” (Jacobs et al., 1981, p. 74). Jacobs’ et al. (1981) writing criteria have content validity since the criteria outlined evaluate the performance of EFL/ESL students’ different types of expository writing which are tasks they perform in the foreign language classroom. The Jacobs et al. (1981) ESL Composition Profile criteria was selected among other similar ones in the present study as it has been used successfully in evaluating the essay writing proficiency levels of students in ESL/EFL programs comparable with those at the Lebanese American University. The Profile is divided into five major writing components: content, organization, vocabulary, language, and mechanics with each one having four rating levels of very poor, poor to fair, average to good, and very good to excellent. Each component and level has clear descriptors of the writing proficiency for that particular level as well as a numerical scale. For example, very good to excellent content has a minimum rating of 27 and a maximum of 30 indicating essay writing which is “knowledgeable — substantive — thorough development of thesis — relevant to assigned topic”, while very poor content has a minimum of 13 and a maximum of 16 indicating essay writing that “does not show knowledge of subject — non-substantive—not pertinent — or not enough to evaluate” (Jacobs et al., 1981). The range for each of the writing skills are content 13–30, organization 7–20, vocabulary 7–20, language 5–25 and mechanics 2–5. Benchmark essays are included in Jacobs et al. (1981) text as guides for teachers. Hamp-Lyons (1990) comments that it is “the best-known scoring procedure for ESL writing at the present time” (p. 78).

Construct validity, the second but equally important issue, is the degree to which the scoring instrument is able to distinguish among abilities in what it sets out to measure, and is usually referred to in theoretical terms; in this case, the theoretical construct is “essay writing ability” which the instrument aims to measure. Jacobs’ et al. (1981) criteria have been researched and found to have construct validity in that significant differences were found when student essay scores were compared.

An additional concern, but of lesser importance compared with the other two issues in this article, is concurrent validity. An evaluation instrument is said to have concurrent validity when the scores obtained on a test using the instrument significantly and positively correlate with scores obtained on another test that also aims to test similar skills. Examples of tests that could correlate with essays might be “...measures of overall English proficiency...” such as the final exam grade in the
English courses even though an essay requires a [similar] writing performance specifically (Jacobs et al., 1981, p. 74) or another sample of writing performance. Jacobs’ et al. (1981) criteria have been established to have concurrent validity; scores being highly correlated with those of the TOEFL and Michigan Test Battery. “The correlation coefficients which result from the relationships between the tests can be considered to be “validity coefficients...” Sixty or above provides strong empirical support for the concurrent validity...”’ for the instrument in question (Jacobs et al., 1981, pp. 74–75). Obviously, in any classroom situation, teachers do not usually compute all these validity tests; however, it is important that they are aware that these issues need to be considered.

Holistic and analytic scoring instruments or rating scales have been used in EFL/ESL programs to identify students’ writing proficiency levels for different purposes. Some instruments may be a combination of both as the Jacobs’ et al. (1981) ESL Composition Profile is. Although in any holistic scoring often specific features of compositions are involved (Cohen and Manion, 1994), holistic scales are mainly used for impressionistic evaluation that could be in the form of a letter grade, a percentage, or a number on a preconceived ordinal scale which corresponds to a set of descriptive criteria (e.g. The Test of Written English as part of the Test of English as a Foreign Language — TOEFL — has a 1–6 scale — Pierce, 1991). Benchmark papers (sample papers drawn from students’ work which represent the different levels) are chosen as guides for the raters, each of whom scores the essays without knowing the score(s) assigned by the other raters (blind evaluation). The final grade is usually the average of the two raters’ scores. If there is a score discrepancy, to be determined by those concerned, a third and sometimes a fourth reader is required and the two or three closest scores or all four scores are averaged. In more explicit holistic scoring, grading criteria are detailed for each of the levels which “…establish the standards for criterion-referenced evaluation…” (Reid, 1993, p. 239) rather than norm-referenced (evaluating students in comparison with each other); then papers can be evaluated across groups. Through rater training and experience, high inter- and intra-reliability correlation can be attained (Myers, 1980; Najimy, 1981; Homburg, 1984; Carlson et al., 1985; Cumming, 1990; Hamp-Lyons, 1990; Reid, 1993; Upshur and Turner, 1995).

Some researchers argue, however, that holistic scoring focuses on what the writer does well rather than on the writer’s specific areas of weakness which is of more importance for decisions concerning promotion (Charney, 1984; Cumming, 1990; Hamp-Lyons, 1990; Reid, 1993; Cohen and Manion, 1994; White, 1994; Elbow, 1999). They do see its value for evaluating classroom essays and large scale ratings such as those done for the Test of Written English international exam, as it saves time and money and at the same is efficient. For student diagnostic purposes, holistic scoring can, therefore, serve to initially identify the students’ writing proficiency level, but for more specific feedback needed in following up on students’ progress and to evaluate students’ proficiency levels for promotional purposes, criterion-referenced evaluation criteria (rather than norm-referenced) are needed which analytic scales provide. A further argument for adopting analytic evaluation scales is that some research indicates learners may not perform the same in each of the
components of the writing skill (Kroll, 1990a), making more qualitative evaluation procedures such as lexical, syntactic, discourse and rhetorical features necessary (Tedick, 1990; Reid, 1993; Connor-Linton, 1995; Tedick and Mathison, 1995).

Analytic scoring scales, in being more criterion-referenced, have also been found to be better suited in evaluating the different aspects of the writing skill. Such scales may also include specific features such as cohesion subsumed under organization [i.e. the structural and lexical signals that connect a text as in Weir’s (1983, 1990) Writing Scale]. One type of analytic scale, primary trait, was used in efforts to obtain more information than holistic scores could provide, to evaluate clearly specific tasks first developed by the National Assessment of Educational Progress (NAEP) in the USA by Lloyd Jones (1977 in Cohen and Manion, 1994). Basically, this scale deals with the setting up of criteria prior to the production of a particular task. Multiple-trait scoring, another analytic scale, focuses on more than one trait in the essay. Often teachers need to consider how students have read, summarized an article or argued one side of an issue (Cohen and Manion, 1994). Teachers need to be aware, however, that in any type of analytic evaluation, they may unknowingly fall back on holistic methods in actual ratings if attention to the specified writing areas is not given. Also, there could be a backwash effect in that instruction is influenced by the evaluation criteria (Cohen and Manion, 1994; Connor-Linton, 1995). Nevertheless, if applied well, these analytic scales can be very informative about the students’ proficiency levels in specific writing areas.

Cohen and Manion (1994) give score results of a few sample essays that were evaluated by the various scales mentioned earlier. They note that there is a variation of results depending upon how raters perform and which scales are used. They conclude that in any writing evaluation training program, raters should focus on the task objectives set, use the same criteria with a common understanding, attempt to have novice raters approximate expert raters in rating, and have all raters sensitive to the writing strategies of students from other languages and cultures (p. 336).

Also, in using the different types of analytic scales, ratings within the specific writing areas may vary from one instructor to the next depending upon the relative weight placed on each. This is quite normal even among the most experienced raters. However, studies have indicated that inter-rater reliability of 0.8 and higher has been obtained (Kaczmarek, 1980; Jacobs et al., 1981; Bamberg, 1982; Perkins, 1983; Hamp-Lyons, 1986a; Bachman, 1990, 1991; Hamp-Lyons, 1991; Alderson and Beretta, 1992; Gamaroff, 2000). Other studies have shown significant correlation between some linguistic features (e.g. vocabulary and syntactic features) as well as between linguistic features and holistic scores (Mullen, 1980).

Analytic evaluation instruments are considered more appropriate to base decisions concerning the extent to which the students are ready to begin more advanced writing courses. Garmaroff (2000) argues that among many factors perhaps why, “...many fail and pass [depends] on who marks one’s tests and exams (who is usually one’s teacher/lecturer); in other words, depending on the luck of the draw” (p. 46). Although this view may be true in some situations, it does seem to undermine what teachers can and should do as part of a team in EFL programs. It is claimed in this article that the overriding factor, whether the student fails or passes
the course is primarily the choice of the scoring scale and criteria suitable for the purpose of the task, and the interpretation (or misinterpretation) of the results. It is also the view of the author that essay evaluation should be team work in any EFL program and not left to individuals alone. This article is an attempt to investigate what two types of writing rating scales, analytic and holistic, can tell teachers in evaluating their students’ final essays. In a sense it is revisiting the essay evaluation scenario and re-raising teachers’ awareness to the significance of their decisions. Often we as teachers are so busy in evaluating piles of essays that a reminder of the significance of knowing what these scales offer us is helpful.

3. Method

3.1. The sample

A stratified random sample of 30 essays was selected from a corpus of $N=156$ final exam essays written by L1 Arabic non-native students at the end of a 4-month semester in the Freshman English I course, the first of four in the EFL program at the Lebanese American University. The final essay (given as part of the final exam which also includes a reading comprehension component) is used to test for promotional purposes. The percentage of the final exam is 30% of the final course grade with the final essay constituting 20% and the reading comprehension 10%. Final grades are reported on the basis of pass (P) or no pass (NP). Students must gain a minimum percentage score of 60% in the course to receive a P. In the final analysis, however, it is the final essay that often determines whether the student is ready to begin the more advanced Freshman English II course. The main objective of these EFL courses is to give students an opportunity to become more proficient in basic reading comprehension and academic essay writing skills.

The sample of essays was written by a representative selection of the student population ($N=156$) in gender (2/3 males and 1/3 female), first language (Arabic 94.55%; French 1.98%; English 1.98%), main study language in high school, either English (35%) or French (60%), (bilingual: English and French 5%), and major in three schools (Arts and Sciences 40.59%; Engineering and Architecture 32.67%; Business 26.73%).

3.2. The procedure

Students were instructed to choose one of two topics, each in a different rhetorical mode, and write a well organized and developed essay within a 90 min time period — considered ample for the task of writing a final essay product (Caudery, 1990; Kroll, 1990b; White, 1994). Although research findings suggest that the rhetorical or organizational mode and topic affect performance (Hoetker, 1982;
Quellmalz et al., 1982; Carlman, 1986; Kegley, 1986; Ruth and Murphy, 1988; Kroll, 1990a; Siegel, 1990; Hamp-Lyons, 1991; Sweedler-Brown, 1992; White, 1994), the topics were considered valid for promotional purposes, as essays were comparable with those given during the regular course instruction. Moreover, the choice gave students the opportunity to perform their best.

Some research has shown that L1 Arabic non-native writers in English show weaknesses when writing in the English script which has been noted to negatively influence raters’ scores (Sweedler-Brown, 1993). However, the essays were not typed before evaluation so as to keep the scoring procedure in a realistic context. Each essay was first given two holistically rated percentage scores using Jacobs’ et al. (1981) ESL Composition Profile according to final course objectives by two readers (the class teacher and another who was teaching a different section of the same course). Using the same profile, the same readers then analytically scored each essay according to the five parts: content, organization, vocabulary, language, and mechanics (Jacobs et al., 1981). The profile had been used in the LAU EFL program by some of the instructors, but this was first time it was rigorously applied.

The final holistic percentage score for each essay was then calculated by recording the mean of the two readers’ scores. In several instances, a third reader was required when discrepancies exceeded one letter-score range (A = 90–100%; B = 80–89%; C = 70–79%; D = 60–69%; Failing = below 60%) in which case the average of the two closest scores were computed. The average mean score obtained for all the 156 essays was 65%, which does reflect the low writing proficiency level expected at this level. The final analytic scores for each of the writing skill components were averaged and a mean for each of the five parts computed.

The Spearman Correlation Coefficient was used to test the strength of the relationship between the two raters’ scores (inter-rater reliability) and between the same raters’ scores on two occasions (intra-rater reliability). This statistical test was considered over the Pearson one since the holistic and analytic scores were not normally distributed. The Friedman’s Statistical Test was used to check for any significant differences on more than two variables; in this case the five specific writing areas in the ESL Profile (SPSS, 1997).

4. Results

Significant positive relations of 0.8 were obtained between the two readers’ holistic essay scores ($P = 0.001$) using the Spearman Correlation Coefficient Statistical Test (inter-rater reliability). When a random sample of essays ($N = 10$) was re-evaluated by the same readers, there was also a similar significant positive relation ($P = 0.001$) between the scores of the first and second readings using the Spearman Correlation Coefficient Statistical Test. That is, raters were in agreement with their own readings of the same essays on two occasions (intra-rater reliability).

Table 1 indicates the mean analytic raw scores and Table 2 the scores when converted to percentages. It is apparent that it is the language component that is the weakest when one compares the percentages.
There were very high positive significant relations of 0.8 and above in the raters’ analytic evaluation \((P < 0.05)\) between the different components of the writing skill.

When the scores of the different components of the writing skill were correlated, Table 3 indicates high significant coefficients, the highest being between the vocabulary and organization components. This might be an indicator that the raters evaluate essays seeing a link between these two components.

All in all, however, the results indicate that the components are inter-related confirming the internal reliability or consistency of the scores (Jacobs et al., 1981).

When the mean percent analytic component scores were compared, there were high significant differences \((P = 0.001)\) using the Friedman two-way ANOVA. That is, students performed significantly differently from best to least as follows:

Content, Organization, Mechanics, Vocabulary, Language

It seems that student performance is the lowest in the vocabulary and language components when compared with that in content, organization and mechanics.

Interestingly, very high significant correlation coefficients \((P = 0.001)\) were obtained between the holistic and the analytic scores when the holistic scores were computed separately with each of the five writing components.

### 5. Discussion

Although there were high inter- and intra-reliability coefficients, the holistic scoring revealed little about the performance of the students in the different components.
of the writing skill. White (1994) warns against holistic scoring alone and suggests that “...we need to define writing more inclusively than our holistic scoring guides...which normally yield no gain scores across an academic year” (p. 266).

However, this finding suggests that the performance of the students is a reliable indicator of the students’ writing proficiency level. That the final essay significantly correlated with the final exam results seems to suggest that writing proficiency is in some way related to academic course work at a given point in time; in this case at the end of the semester (concurrent validity).

Jacobs et al. (1981) point out that an important aspect in testing is the internal consistency or reliability of scores; that is, there should be no significant correlation differences among the parts or between the parts and the whole score (p. 71) if the test is to be considered a reliable one. In the study, there was high significant correlation between the two raters’ analytic scores on each of the five components as well as among the scores of the different components of the writing skill, the highest being between vocabulary and organization. Although based on one writing sample, this reinforces some recent research (Hoey, 1991) that claims that lexis and organization are related in various ways over stretches of discourse.

When the analytic scores were compared with each other, there were high significant differences among the different writing components. That is, students performed significantly differently in the various aspects of the writing skill. This finding confirms research in the field (Kroll, 1990a) that students may have different proficiency levels in the various writing components. EFL programs and the teachers should re-visit the value of the scoring instruments adopted and consider not relying entirely on holistic scores, if this be the case, to determine the improvement of their students’ writing proficiency. Decisions concerning students passing or failing the course would be better informed ones if not totally based on the holistic rated essay scores. There have been many experiences where students have been passed due to their having ‘relevant’ ideas and ‘appropriate’ organization, but their relatively poor language and vocabulary repertoire have placed them at a disadvantage in these upper level English courses. Of course, the reverse has also been the case. The fact that the language component indicated the lowest scores as identified by the criteria confirms both teacher and student perceptions during informal interviews throughout the course. This is quite a significant finding (not revealed in the holistic rating) as it implies that incoming students need to work on widening their vocabulary repertoire in an academic setting; a finding which confirms much of

Table 3
Correlation among analytic mean percent scores

<table>
<thead>
<tr>
<th>Component</th>
<th>Organization</th>
<th>Vocabulary</th>
<th>Language</th>
<th>Mechanics</th>
<th>Content</th>
<th>Organization</th>
<th>Vocabulary</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organization</td>
<td>0.8774</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vocabulary</td>
<td>0.8962</td>
<td>0.9587</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Language</td>
<td>0.7945</td>
<td>0.8149</td>
<td>0.8679</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mechanics</td>
<td>0.6871</td>
<td>0.7607</td>
<td>0.7712</td>
<td>0.7751</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Content</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.8774</td>
<td>0.9587</td>
<td></td>
</tr>
</tbody>
</table>

\( P = 0.001 \) in all cases. \( N = 30. \)
the literature in the field, specifically related to L2 Arabic speakers of English (Doushaq, 1986; Silva, 1993).

There is a need to look deeper into the students’ language component to describe what actually is going on and devise ways to help students widen their vocabulary repertoire. The fact that the content component mean percentage score is significantly higher in comparison to the other components indicates that the students had relevant ideas on the topic given, but they found it significantly more difficult to organize and express their ideas. This finding also confirms the teachers’ experiences and students’ comments throughout the course during informal interviews.

However, it might be argued that each component in the ESL Composition Profile (Jacobs et al., 1981) is weighted differently, which might have affected the raters’ scores; that is raters may have been influenced by the score values for each component of the compositions, and implicitly given higher scores to aspects of the writing that the rating scale gives higher weight to (e.g. content) and correspondingly lower scores to the aspects of writing that the rating scale gives lower weight to (e.g. mechanics). Also, to compare these scores with one another more validly, they may have needed to be converted to (or better, scored on) a common scale (e.g. using z scores or some other type of conversion). However, since these components are weighed differently in the teaching and learning process as expressed in the profile, the scale was found to be appropriate for the present study.

The fact that there were high significant positive relations ($P = 0. < 0.001$) between the holistic and the analytic scores using the Spearman Correlation Coefficient Test indicates internal consistency or reliability of the scores. It seems that the readers were focusing on each of the components equally when rating the essays and did not emphasize any component over the other as some of the research suggests might happen. For example, Shakir (1991) reported findings where there was a much higher correlation between the grammar score in comparison with other components and the coherence holistic scores. The research concludes that teachers may have been rating the essays according to more surface features at the sentence level rather than discourse features over larger stretches of text. Sweedler-Brown (1992) also reported high significant correlation between sentence features (rather than more rhetorical ones) and holistic scores and recommends that experienced writing instructors need to work with teachers that are not trained in ESL/EFL evaluation techniques. The consistent high significance of the relation between the vocabulary and the holistic essay ratings may suggest that teachers consider vocabulary a very important sub-component of the writing skill, whereas mechanics is not as important in holistic ratings.

6. Conclusion

The aim of the present study was to find out what holistic and analytic evaluation can tell us and what general lessons can be drawn for the evaluation of writing. Although holistic scoring may blend together many of the traits assessed separately in analytic scoring, making it relatively easier and reliable, it is not as informative
for the learning situation as analytic scoring. Although the study was done on a limited sample, the results indicate that more attention should be given to the language and vocabulary aspects of students’ essay writing and a combination of holistic and analytic evaluation is needed to better evaluate students’ essay writing proficiency at the end of a course of study. In the final analysis, relevant evaluation criteria go hand in hand with the purpose upon which the criteria, benchmark essays and training sessions are based (Pierce, 1991; Elbow, 1999). Most of all perhaps, these initial results confirm the complexity involved in choosing rating scales and delineating criteria for valid and reliable essay evaluation on which to base promotion decisions. What holistic and analytic evaluation can tell teachers about our students’ essay writing proficiency is quite a lot for any EFL/ESL program in any context; the choice of instrument and how we use the information obtained, depends very much upon what we in evaluating our students’ work are looking for.

References


Bachman, L., 1991. What does language testing have to offer. TESOL Quarterly 25 (4), 671–672.


Perkins, K., 1983. On the use of composition scoring techniques, objective measures and objective tests to evaluate ESL writing ability. TESOL Quarterly 17 (4), 651–671.


Raymond, J.C., 1982. What we don’t know about the evaluation of writing. College Composition and Communication 33 (4), 399–403.


Ruth, L., Murphy, S., 1988. Designing writing tasks for the assessment of writing. Ablex, Norwood, NJ.


