Research Issues

Evaluation of ESL Student Writing on Text-responsible and Non-text Responsible Writing Tasks [1]

Heather Boldt, Maria Ines Valsecchi and Sara Cushing Weigle,
Georgia State University [2]

Background

Evaluation of ESL student writing on text-responsible and non-text respon-
sible The issue of how English as a Second Language (ESL) and English as a Foreign
Language (EFL) student writing is perceived and evaluated by instructors, both in
language programs and in various academic disciplines of the universities, is a very
important one for several reasons. First of all, professors' decisions have an impact
on students' academic careers, and thus it is important to understand the extent to
which features of ESL writing play a role in how writing in content areas is evalu-
ated. Secondly, any differences revealed in how ESL teachers and content-area pro-
fessors judge ESL student writing would have implications for any program in the
United States or abroad that wishes to successfully prepare ESL students for univer-
sity content courses in the United States. In other words, ESL and EFL curriculum
developers and program administrators need to know how to assess students for
placement into English programs and to know what is expected of ESL students
when they enter United States universities as mainstream students. As pointed out
by Vann, Meyer and Lorenz (1984), "finding out how [faculty outside of the pro-
gram] view typical ESL writing errors is an important step in developing priorities
for ESL composition instruction" (p. 427).

---

[1] This is a refereed article.

[2] The authors can be reached at thier email addresses: ESLSCW@langate.gsu.edu ; heatherboldt@yahoo.com; mivalsecchi@arnet.com.ar

A considerable amount of research has endeavored to establish how content and ESL faculty make judgments about ESL student writing. Cumming (1985), for example, in his study of ten ESL instructors' responses to ESL writing, found that the teachers differed in their attention to syntax and content. And in Mendelsohn and Cumming's (1987) analysis of ESL, Engineering and English professors' weighting of rhetorical organization, language use and accuracy, the researchers found that the content professors and ESL professors differed in the relative importance that they placed on language use and rhetorical organization. While the content professors (especially the Engineering professors) placed more emphasis on language use, the ESL professors said that rhetorical organization was the more important factor. In a 1988 research study investigating the relative severity of Social Science and Physical Science professors' judgments of non-native speaker (NNS) academic writing, Santos found that professors in both academic disciplines judged content more severely than language. Although the Physical Science professors were not as lenient as Social Science professors were, they did not rate the content lower in papers with language errors, suggesting that these professors consider content and language independently.

One limitation of the studies cited above is that the writing that was evaluated by content professors was restricted to general topic essays of the type that are typically used on English proficiency examinations or written in many ESL and EFL classes. As Leki and Carson (1997) point out, however, most academic writing for content courses outside of English composition is "text-responsible"; that is, students are required to demonstrate their understanding of specific texts (broadly defined as reading texts, lectures, or other sources of information) in their writing, rather than writing from personal experience or using a source text as a springboard for their own ideas and opinions. Little is known about how content area professors judge "text-responsible" ESL student writing and to what extent accuracy of content versus linguistic accuracy affects these judgments. This is an important issue for English for Academic Purposes (EAP) programs, since the goal of such programs is to prepare students for academic coursework in content areas.

The study described in this paper is a pilot study intended to investigate the differences both in performance and in evaluation of text-responsible writing and non-text-responsible writing by ESL instructo ic and content-area professors. Specifically, the research questions are the following:

(a) Do content professors and ESL instructors evaluate ESL student writing differently?
(b) What features of writing do ESL and content professors attend to when rating ESL student writing on text-responsible and non-text responsible writing tasks?
(c) What effect does subject-area knowledge have on writing assessment? That is, will content area professors be more or less lenient on writing within their own field than in other fields?

Because this is a pilot study intended to generate additional hypotheses and research questions, we collected data from a small number of students and raters. We believe that an intensive look at the data from a variety of perspectives can provide guidance for ourselves and other researchers, which will be useful in stimulating additional research in this area.

Method

Participants

A total of six faculty members from Georgia State University (GSU) participated in this study (see Table 1). They belonged to the departments of History (2), Psychology (2) and Applied Linguistics and ESL (2). The age of the raters was between 30 and 60; two were male, and four were female. All of them were native speakers of English, and had had some previous experience in working with both native and non-native speakers of the English language. One rater was competent in a foreign language. Each rater read and provided a holistic score of eighteen NNS compositions written by students in the Intensive English Program at GSU, filled out a background questionnaire, and completed a scoring sheet attached to each composition they read. (See the table on the next page)

Materials

The background questionnaire (adapted from Santos, 1988; see Appendix 1) requested the following information from each rater: name, department, gender, age, native language, knowledge of a second language, teaching experience in the field, number of native speakers (NS) in the professor's classes, approximate number of non-native speakers (NNS) the professor had taught, whether these students tended to be graduates or undergraduates, and the professor's policy in dealing with both NS and NNS writing errors. For this purpose, a multiple choice option was provided for which the participants had to select the one that reflected their scoring tendency: a- does not correct or downgrade language errors, b- does not correct language errors but downgrades for them, c- corrects language errors but does not downgrade for them, and d- corrects and downgrades for language errors.

After reading each composition, the raters were also asked to fill out a short scoring sheet. This scoring sheet was slightly different for content and ESL faculty (see Appendix 2 and 3, respectively). Raters were first asked to evaluate if the writer of the composition had the English language skills necessary to be successful at college-level work. A multiple-choice selection was provided (a- *yes, definitely*, b- *possibly*, c- *probably not*, and d- *definitely not)*. The second item in the scoring sheet asked raters to rank order the features they thought had been important in making the previous decision. The three categories provided were content, gram-

Rater Code	*Sex*	*Age*	*Native Language*	*Other Languages*	*Teaching Experience*	*Percentage of NNS*	*Grad/ Undergrad*	*# of NNS*	*NS Policy**	*NNS Policy**
H1 (History 1)	M	40-49	English	French	7-9 Ys	15	Undergrad	500+	3	3
H2 (History 2)	F	40-49	English	None	4-6	15	Graduate/Undergrad	300+	4	4
P1 (Psychology 1)	F	30-39	English	None	4-6	5	Undergrad	10	3	3
P2 (Psychology 2)	M	30-39	English	None	10-12	5	Undergrad	20	3	3
E1 (ESL 1)	F	30-39	English	None	7-9	n/a	Undergrad	1200+	n/a	4
E2 (ESL 2) *(sometimes 3)*	F	50-59	English	None	13+	10	Undergrad/	250+	1	1

* The policy describes the way the professor generally deals with student writing:

- 1- does not correct or downgrade language errors
- 2- does not correct language errors but downgrades for them
- 3- corrects language errors but does not downgrade for them
- 4- corrects and downgrades for language errors.

Table 1. Rater profiles

mar and organization. The last question in the sheet was related to whether content professors thought that this particular student could pass an introductory course in his/her academic field. Content professors were given these three questions while ESL instructors were only given questions number 1 and 2.

The compositions *(for the purposes of this paper, the terms "composition" and "essay" are being used interchangeably to refer to the writing samples providd by students, which ranged from a single paragraph to a two-page essay)* that raters scored belonged to six non-native speaker (NNS) students studying at the Intensive English Program (IEP) at Georgia State University. These students were in their last semester in the IEP; students at this level generally have TOEFL scores between 500-525 and may already be enrolled in one or two mainstream academic courses. Students whose essays were used for the study were enrolled simultaneously in two of the IEP's content-based courses: Advanced Reading and Listening, which uses psychology as content, and Advanced Writing for University Exams, which has U.S. history as its content. The writings with history and psychology content were taken from tests provided to us by the IEP course instructors. In addition, we administered a 30-minute timed writing test using a disclosed prompt from the TOEFL (ETS, 1999) (see Appendix 4) to collect a third writing sample from these same students. Complete data were available for six students who were present on test days in both courses and who also were present on the day that the general writing prompt was administered.

The general essays were responses to a question about learning alone versus learning with a teacher. Since this writing task does not require any specific content, it is considered to be a non-text-responsible task. Because the purpose of the tests in the IEP courses was for students to demonstrate their mastery of course content, the psychology and history writing tasks are considered text-responsible. The psychology essays were short essay (paragraph-length) responses to a question about desegregation. The history essays were responses to two different essay questions because the six students were enrolled in two different sections of the Academic Writing course; one question involved explaining why a certain person was important in US history and the other question involved comparing and contrasting the early history of two states. In contrast to the paragraph-length psychology essays, the history essays were generally about a page in length.

Procedures

As stated above, six students provided data for the study, resulting in a total of 18 compositions (see Table 2 for demographic information about the students). These 18 compositions were then labeled with an identification letter so that the raters would not know that three of the compositions belonged to the same students.

Participant	Gender	Native Language
Student 1	Male	Urdu
Student 2	Male	Hindi
Student 3	Female	Lithuanian
Student 4	Male	Tagalog
Student 5	Female	Vietnamese
Student 6	Female	Nepali

Table 2. Student profiles

The compositions were then sent to the six raters who had volunteered for this study. A cover letter was included, which indicated that the papers were paragraphs extracted from tests administered in IEP courses, and instructed the raters to complete the questionnaire, read the 18 compositions, fill out the scoring sheet attached to each paragraph, and return them.

Analysis & Results

Procedures for data analysis and results are presented separately for each of the three research questions.

Research Question #1:
Do content and ESL professors rate NNS writing tasks differently?

Analysis

We calculated the total average given to the 18 student essays by assigning a numerical value to each of the raters' answers to the question "Do you think this student has the English language skills to be successful at college-level work?" The numerical values we assigned are outlined below:

4 = yes, definitely
3 = possibly
2 = probably not
1= definitely not

Results and Discussion

Using these values, the overall average for the 18 student essays was 3.06, with individual averages ranging from 2.50 for rater P1 to 3.68 for rater H1 (see Table 3). Since there were only two raters in each discipline, comparisons between raters in each discipline have limited meaning, but as the table shows, the history professors as a group were somewhat more lenient than the psychology professors,

with the ESL instructors in the middle. These results show that there was, in essence, no difference in the overall ratings of the content professors as a group and ESL professors.

	Individual raters						Raters grouped by discipline			
Rater	H1	H2	P1	P2	E1	E2	History	Psych	ESL	Total
Mean	3.68	2.94	2.50	3.11	3.33	2.78	3.31	2.81	3.06	3.06
s.d.	0.59	.80	.71	.83	.69	1.00	.79	.82	.90	.85

Table 3. Average scores given by raters (number of ratings per rater=18)

We had hypothesized that ESL professors might be more lenient than content professors because of their familiarity with ESL writing processes and difficulties, but this does not seem to be the case with these particular raters. Further research with a larger pool of raters is needed, but this result is similar to Brown's (1991) finding that the scoring of ESL and English professors did not vary significantly when holistically scoring 112 student essays.

Although the raters from the various faculties came to the same conclusions, we began to wonder if there were some underlying differences in the way ESL and content professors approached the rating task. To borrow Brown's (1991) words, the "faculties may have arrived at their scores from somewhat different perspectives" (p. 567). For a detailed look into this possibility, we had to consider the data in relation to our second research question.

Research Question 2:
What features of the writing sample do ESL and content professors attend to when rating NNS writing on text-responsible and non-text responsible writing tasks?

Analysis

For this question, we analyzed the results on the scoring profile that asked the raters to "rank the following three aspects in order of how important they were in making the above [rating] decision." The three aspects to rank order were content, grammar and organization. To get an overall view of which aspects played the most significant role in evaluation, we tallied the number of times each professor had chosen content, organization or grammar as their first priority, i.e., how many times did Professor X list content as the most important factor? Organization? Grammar? In Figure 1 below, as in the other figures to follow, H1= History Professor 1, H2= History Professor 2, P1= Psychology Professor 1, P2= Psychology Professor 2, E1= ESL Professor 1 and E2= ESL Professor 2.

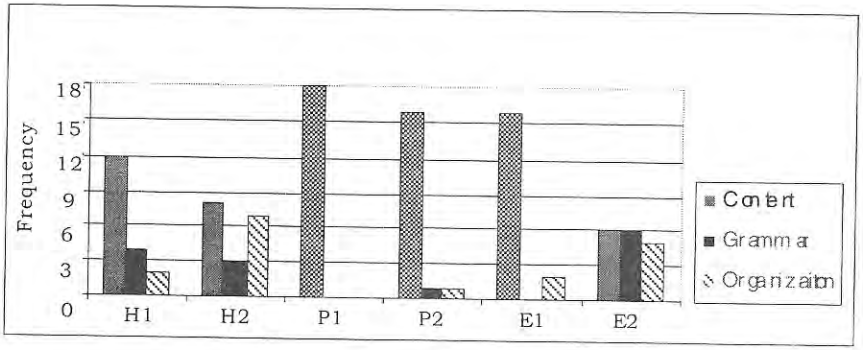

Figure 1. First ranking of each rater

Since we were interested in patterns of rating within academic disciplines, we then combined the information from Figure 1 into three categories: History, Psychology and ESL (see Figure 2).

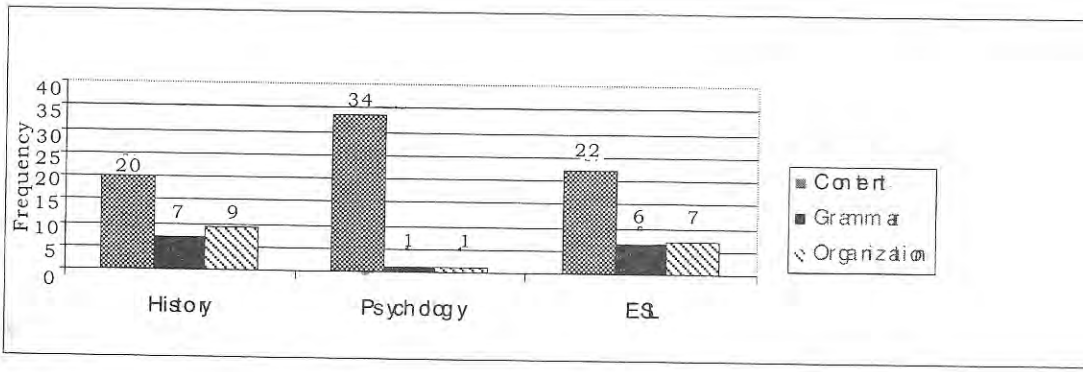

Figure 2. First ranking of raters, by discipline

Results and Discussion

As can be seen in Figure 2 above, raters from the three departments considered content as their priority, with grammar and organization playing a secondary role. In fact, when considered together, psychology professors almost always attended to content (only in two instances did psychology faculty list something other than content as most important), while History and ESL faculty showed a similar pattern (choosing something other than content first numerous times).

Having established that content was the most important factor for all raters, the next step was to investigate the relative importance of grammar and organization in raters' evaluations. In other words, what is the next most important factor after content? Table 4 below shows the rank orderings of content (C), organization (O) and grammar (G), categorized by the rating of 4-1 given by the raters (4 = definitely could pass, 1 = definitely not). For example, on 24 occasions raters gave a score of 4 and reported a rank ordering of COG (content followed by organization followed by grammar) while there was only one time that a paper received a 4 with a rank ordering of GOC (grammar followed by organization followed by content).

	Rank Ordering of Grammar, Content & Organization						
Rating (4-1)	**COG**	**CGO**	**OCG**	**OGC**	**GCO**	**GOC**	**Total**
Definitely pass (4)	24	2	2	4	2	1	35
Possibly pass (3)	24	7	5	2	2	4	44
Probably not (2)	6	9	1		1	2	19
Definitely not (1)	2			1	1		4
Essay total	56	18	8	7	6	7	102*

**Total is not 108 because rater E2 did not provide rank orderings for 6 essays.*

Table 4. Rank ordering of content, organization, and grammar categorized by overall rating the essay received

As can be seen by looking at the OCG, OGC, GCO and GOC columns, these particular rankings were used infrequently and sporadically across all levels of papers; no pattern can be found here. In the first two columns of the table, however, in with content is listed as most important followed by either grammar or organization, it does seem as if overall writing quality may be playing a role in the feature that raters attend to after considering content. Although it is hard to generalize from such a small sample, the ranking pattern of COG, which was found in 54% of the papers across all raters in all subject areas, is most evident in the ratings of papers that received high values (3 and 4). For the lower values (1 and 2), however, the COG pattern is not nearly as predominant. In fact, it seems that when a paragraph is not perceived as very successful, it is equally likely to be due to deficiencies

in organization as to deficiencies in sentence-level grammar; that is, the chances o grammar being considered second most important is just likely as organization (8 out of 17 instances versus 9 out of 17 instances respectively).

To summarize, when all three writing tasks are considered together, our data shows that raters attend first and foremost to content. With strong writing (here papers that received 3s or 4s), raters then consider organization followed by grammar. With weaker writing, however, once content has been considered, raters show no strong tendency to either organization or grammar.

Task Effects on Ratings

Since students each wrote two text-responsible essays (in psychology and history) and one non-text-responsible essay (the general writing prompt) we wondered whether there was a relationship between the writing task itself and raters' ordering of the three aspects of writing (content, organization, and grammar). Table 5 shows the same ratings as Table 4, categorized this time in terms of the writing task rather than the score that was given. As the table shows, the COG pattern was particularly predominant for the short-answer psychology essays, with 26 out of 32 ratings. The COG pattern was also the dominant pattern for history essays, although there was more variation in patterns with these essays. In particular, organization was selected as the first consideration eight times for these essays. In contrast to the text-responsible essays, ratings on the general essays were much more grammar oriented: while content was still the first consideration for over half of the essays, the most prevalent pattern was CGO rather than COG, and raters chose grammar as the first consideration in 10 out of 35 ratings.

	Rank Ordering of Grammar, Content & Organization						
Content	**COG**	**CGO**	**OCG**	**OGC**	**GCO**	**GOC**	**Total**
Psychology	26	4	1		1		32
History	22	2	6	3		2	35
General	8	12	1	4	5	5	35
EssayTotal	56	18	8	7	6	7	102*

**Total is not 108 because rater E2 did not provide rank orderings for 6 essays.*

Table 5. Ranking of content, organization, and grammar categorized by content area of essay

To summarize, for the psychology essays, ratings were overwhelmingly focused on content, with organization of secondary concern and grammar of minimal importance. For some of the history essays, organization became a more important factor in raters' decisions, and for the general essays, grammar was more of a focus

There are several possible explanations for these results. In the case of the psychology essays, the essays tended to be very short (a paragraph at most), and the subject matter was familiar enough to raters in the American context so that raters from all disciplines could discern whether the content was accurate or not. For example, the excerpt below is from an essay from Student 6; it received passing scores (3 or 4) from five of the six raters:

> Desegregation, first mix all students in same school, and their curriculums are same, so the African American students can get same amount of attention, facilities, same test matters also, they have chance to improve their academic achievement.

In contrast, the following excerpt is from an essay from Student 3 that was given scores of 1 or 2 by five of the six raters:

> Statistics say that African Americans do worse in shools that white American students. That is because, of desegregation:
> * blacks have they own believes about the schools.
> * blacks have different morality about the schools.

Even though four of the six raters were not content-area experts, the content of this particular test item allowed raters to discern whether the content was accurate. While both excerpts have grammatical problems, the content of the first essay clearly addresses the test question more accurately than that of the second essay.

In contrast, the history essays dealt with facts and details that raters may not have been familiar with, and required students to provide substantially more information than the psychology task. In this case, the organization may have played a more important part in the rating, since the accuracy of the content was not always obvious to raters. Furthermore, if the facts are presented in the essay but the organization is unclear, raters may have difficulty understanding the content. The essay from Student 5 below is an illustration of the importance of effective organization:

> ***Prompt:*** *Explain why Anne Hutchinson was important in US History*
>
> ***Student Response:*** *According to the text, Anne Hutchinson was important in US History with three reasons. The first reasons was her belifes. Anne Hutchinson belived that every people could have a direction conversation with God and they would be save by God directly _ "covenant of grace". In addition, Anne Hutchinson also mention about "covenant of good words" which means if som eone who did a good work, he/she also had God's salvation. Moreover, the second reasons that made Anne Hutchinson was important in US History was because of she was a puritan. Her belifes was made the Puritan angry because of she against them. And the last reason because of Anne Hutchinson was a women. A women could not saying something have much that power at that century.*

> *Therefore, all the puritan men was not accept her saying and her against them like that. Because Anne Hutchinson had all these three reasons, she was important in US History.*

This essay received passing scores from all six raters, despite a large number sentence-level errors. It seems as if the student's organization in the history pap compensated for grammatical weakness. The overall structure of the essay is cle and the transitions are used well; this clear organization of facts led to a passi grade by all six raters. With Student 1, on the other hand, we have an example of poorly organized essay that resulted in the only two failing grades given by a raters to history essays:

> ***Prompt:*** *Compare and contrast the early years of Maryland and Virginia*
> ***Student Response:*** *Ther were lot's of similarities and difference's between Virginia and Maryland. One of the Similarity was that both were founded by English Settlers in 1600's. The leaders brought servats with them. Secondly both colonizations grew several other products but mostly depended on tobacco. Since they had Similarities but both shared some differences too. Likewise Virginia was granted to the Virginia Company and they only had the authorities to pass laws in Virg. This company wanted to make money from their stock holders. They were looking for Gold and other valuable and the passage to Asia. On the other hand, Maryland was granted to Lord Baltimore Sir George Calvest who and the landowners had the authoity to make rules and regulations for there colonization. Lord Baltimore wanted mostly Catholic's who were discriminated in England, and all other refuge's. In Virginia they were going to give lands to those people who cold bring 50 men's more to increase their population.*

Even though most of the facts needed to answer the question are provided in t above essay, they are hard to find because of the misused transitions and lack cohesion. While content is the feature most important to raters, these results imp that the organization of the facts can make or break an essay, particularly in cas where the readers may not be familiar with the content. With student #5's histo essay, the good organization compensates for poor grammar and not very stro content. With student #1's history essay, we have an example of poor organizati outweighing the content, even though it is very strong, because it is difficult follow.

We now turn to the analysis of research question number three.

seeseemed to be stricter on the non-text-responsible essays than the facul

Research Question #3
What effect does subject-area knowledge have on assessment? Will content professors be more or less lenient on paragraphs within their subject area?

We had hypthesized that content professors would be more severe in the rating of students paragraphs in their own field. To address this question, we calculated the number of times that raters gave passing scores (i.e.,scores of 3 or 4) to essays in each field. The results of this analysis are found in Figure 3.

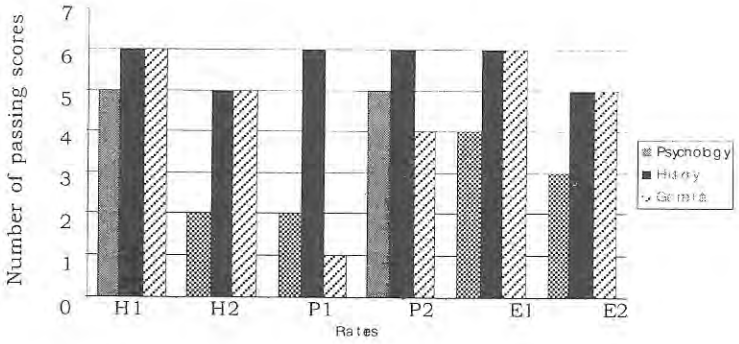

Figure 3. Number of passing scores given by individual raters to essays in different disciplines

Results and Discussion

As the figure shows, the history essays received the most passing scores, with all raters passing at least five essays. The psychology essays, on the other hand, received the worst scores, with most raters giving more failing scores to psychology essays than to either history or general essays. One interesting finding is that the ESL and history professors seemed to have similar grading standards for the history and general essays, with the same number of passing scores given to each group of essays by each of these raters. The psychology professors, on the other hand, rated the general essays lower than the history essays: P1 in particular gave only one general essay a passing score. While the sample is too small to make any generalizations about these results, there does not seem to be a clear pattern of leniency or strictness within fields: most raters gave higher scores to the history essays than to the psychology essays. On the other hand, the psychology professors

seemed to be stricter on the non-text-responsible essays than the faculty from other fields. This result may be due to the data-driven nature of psychology, where clear presentation of concrete facts may be valued over the originality of ideas and interpretations, as in the humanities. In short, the differences between professors in different fields shows up not in terms of the professors' own fields versus other fields, but rather in terms of text-responsible essays where clear presentation of concrete facts may be valued over the originality of ideas and interpretacion, as in the humanities. In short, the differeces between professors in responsible essays where clear presentation of facts is crucial, versus non-text responsible essays, where the arrangement and development of ideas is more important. It would be difficult to generalize beyond this particular sample, but these results suggest that a fruitful area of inquiry would be to gather data from more faculty in different disciplines on ESL text-responsible versus non-text-responsible writing.

Discussion

After the analysis of our data, we can conclude that no significant difference was found between the way content and ESL professors rate NNS compositions in terms of their overall scores, when all the writing samples were considered together. Our hypothesis that ESL raters might be more lenient, at least with the small number of participants we had for this study, was rejected. We think that the proximity in means between the two groups of raters is positive for a number of reasons. First, it means that the ESL professors are realistic with what happens outside the ESL department, and that their way of rating is in tune with the rating carried out by professors in other departments. Secondly, it shows ESL students that they are receiving adequate preparation for entering the university, where they are not likely to find many discrepancies in the way their writing will be viewed and scored. Third, this balanced result gives the IEP at Georgia State University encouragement and confidence in the way it assesses students' performance level. Content professors from the History and Psychology department agreed with the ESL professors that overall these students had the language skills necessary to go into university life. This shared judgment can give ESL instructors confidence in the way they are approaching the teaching and assessment process of the fifth and last level of the IEP.

As for what aspects of the writing most professors attended to first when rating, the answer in favor to content was overwhelming. In 71 percent of the cases, content was rank-ordered first. However, a clear second place could not be identified. Organization and grammar were rank-ordered first in about the same number of cases. With the rank ordering of Content, Organization and Grammar, it is interesting to notice that the pattern Content-Organization-Grammar was most predominant with the high quality compositions (those scoring 3s and 4s). However, with the lower-quality compositions (scoring 2s and 1s) a systematic pattern could not be found (in other words, with these compositions, the patterns COG and CGO appeared about the same number of times. Our hypothesis, in this respect, is that compositions that are rated highly have clear content and smooth organization;

teachers do not need to pay much attention to grammar because the message has been successfully conveyed through the other two aspects. On the contrary, raters tend to give lower scores to compositions with unclear content, and in such cases it may be difficult to determine whether grammar or organization is interfering the most with the clarity of the content. This could be the reason that grammar and organization show up as the second choice about the same number of times. We could think of two possible scenarios that would account for the fact that we see grammar as the second most important factor just as often as we see organization as the second most important factor:

1) The rater, because the content is lacking and there seems to be no organization, does not take the time to consider the grammar (i.e. organization is the second factor in decision making);
2) The rater, because the content is lacking and the grammar is hard to understand or is perhaps very irritating, does not even begin to consider the organization (i.e., grammar is the second factor in decision making).

Although we did not ask the raters to provide reasons for their choices, this second possibility seems plausible in light of Brown's (1991) study which found that grammar was a primary **negative** feature in rating.

Another note regarding research question two is that the patterns of ranking of content, organization or grammar as the most important aspect in evaluation were very similar for history and ESL professors (refer again to Figure 3). Both ESL and history professors listed grammar and organization as most important about the same number of times (between 13 and 18 times), while psychology professors only listed grammar and organization as the most important factor one time each. The strong tendency towards attending to content that was displayed by the psychology professors may be due to the data-driven nature of the field. History and ESL, however, are in the humanities and more concerned with writing and as such professors in these fields may be more concerned with the role that grammar and organization should play in good quality writing.

Finally, the results of the analysis of task effects on ratings showed that grammar played a more important role in the ratings for the non-text responsible essays than it did for the text-responsible history and psychology essays. This result has implications for using non-text responsible essay prompts in proficiency examinations: specifically, if further research in this area confirms the findings presented in this paper, the practice of using general writing prompts as an indicator of academic writing ability in content courses may be called into question, since different criteria may be used to judge text-responsible and non-text responsible essays.

Regarding our third research question, we cannot provide a definite answer about whether subject-area knowledge had a role to play in the way content profes-

sors rated their area-specific compositions, making them more or less severe. All raters regardless of discipline tended to score the history essays higher than the psychology essays. However, the psychology professors rated the general essays more strictly than did the ESL or history instructors. We believe the few number of raters participating in this study makes it impossible to generalize beyond the current study. In future research, the inclusion of a greater number of content raters could provide more generalizable findings.

Suggestions for future research

The research presented in this paper is a first step towards increasing our understanding of how ESL student writing is evaluated by content professors. While the results that have been presented here cannot provide any definitive answers, we believe that they provide some intriguing possibilities for further research in this area. Of course, we feel that the number of raters participating in a future study should be increased and that raters outside the United States participate. This would provide more reliable and generalizable results. Similarly, future studies should include a wider variety of writing tasks from students at different levels of language proficiency.

As a complement to the quantitative data collected for this study, we believe it would be useful to include raters' think-aloud protocols in order to get richer information as to why certain decisions were made, what factors influenced those decisions, and whether the feature of writing they rank-ordered first was more important because it was weak or because it was strong. We also suggest further research on the writing task's effect on salient characteristics. Our research indicates that several factors related to the nature of the task (text-responsible versus non-text responsible writing, short answer versus essay, social science versus humanities content) may play a role in the way raters rank order content, organization and grammar. More research is needed to separate out these different facets of the writing task.

Conclusion

As Leki and Carson (1997) point out, writing text-responsible essays for content courses creates both challenges and opportunities for ESL students. It is important for ESL and EFL teachers preparing students for academic work in the United States to understand how this kind of writing differs from the non-text-responsible writing that is typical of many writing classes, and it is equally important to understand how students' writing will be evaluated by their professors once they begin their university studies. While the study presented in this paper does not provide any definitive results, and perhaps raises more questions than answers, we hope that the work presented in this paper will be used as a starting point for others who wish to pursue research in this important area.

Acknowledgements

We would like to thank the six raters who participated in the study and the ESL instructors who allowed us to collect data in their classrooms for their time and effort. We would also like to thank John Bunting, Lisa diPuma, Joan Carson, and two anonymous MEXTESOL Journal reviewers for their helpful comments on earlier drafts of this paper, and Dana Maslekoff for her editorial assistance.

[1] For the purposes of this paper, the terms "composition" and "essay" are being used interchangeably to refer to the writing samples provided by students, which ranged from a single paragraph to a two-page essay.

References

Brown, J.D. (1991). Do English and ESL faculties rate writing samples differently? TESOL Quarterly, 25, 587-603.

Cumming, A. H. (1985). Responding to the writing of ESL students. In A. Pare and M. Maguire (Eds.). Patterns of Development. Ottawa: Canadian Council of Teachers of English.

Ferris, D. & Hedgcock, J.S. (1998). Teaching ESL composition. Mahwah, NJ: Lawrence Erlbaum Associates.

Leki, I. & Carson, J. (1997). Completely different worlds: EAP and the writing experiences of ESL students in university courses. TESOL Quarterly, 31, 39-69.

Mendelsohn, D. & Cumming, A.H. (1987). Professors' ratings of language use and rhetorical organizations in ESL compositions. TESL Canada Journal, 5, 9-26.

Santos, T. (1988). Professors' reactions to the academic writing of non-native speaking students. TESOL Quarterly, 22, 69-90.

Silva, T. (1993). Toward an understanding of the distinct nature of L2 writing: The ESL research and its implications. TESOL Quarterly, 27, 657-677.

Sweedler-Brown, C.O. (1993). ESL essay evaluation: The influence of sentence-level and rhetorical features. Journal of Second Language Writing, 2, 3-17.

Truscott, J. (1996). The case against grammar corrections in L2 writing classes. Language Learning, 46, 327-369.

Vann, R.J., Meyer, D.E. & Lorenz, F.O. (1984). Error gravity: A study of faculty opinion of ESL errors. TESOL Quarterly, 18, 427-439.

APPENDIX 1 – BACKGROUND QUESTIONNAIRE

Instructions: Please fill out the following questionnaire and return it to us with the student essays that you have rated. We thank you again for your participation.

Name _____
Department _____

What is your gender?

Male Female

What is your age?

20-29 years old30-39 years old 40-49 years old

50-59 years old 60+ years

Are you a native speaker of English?

Yes No

If no, please indicate your native language here: _____

Are you proficient in any other languages?

Yes No

If yes, list language(s) here: _____

How much experience do you have teaching in your field?

1-3 years 4-6 years 7-9 years

10-12 years 13 or more years

What percentage of non-native speakers do you normally encounter in your classes?
(*This question for content professors*)

approx. 1% approx. 5%

approx. 10% approx. 15%

In raw numbers, about how many non-native speakers of English have you taught?

Less than 5 Less than 10
Less than 15 More than 15

What type of students were the non-native speakers? Graduate Undergraduate

Which of the following statements best describes the way you deal with **native speakers** of English?

I do not correct or downgrade (i.e. lower the grade) for language errors
I do not correct language errors but I downgrade for them
I correct language errors but I do not downgrade for them
I correct and downgrade language errors

Which of the following statements best describes the way you deal with **non-native speakers** of English?

I do not correct or downgrade (i.e. lower the grade) for language errors
I do not correct language errors but I downgrade for them
I correct language errors but I do not downgrade for them
I correct and downgrade language errors

Adapted from Santos (1988)

Appendix 2 – Scoring Profile for Content Professors

Rater's Initials: _____

Scoring Profile (Content Professors)

Instructions: Please fill out the form after reading the attached paragraph. On some of the attached sheets, there is more than one paragraph. Please answer the following questions for the highlighted paragraph only.

1) Do you think this writer has the English language skills to be successful at college-level work?

Yes, definitely
Possibly
Possibly not
Definitely not

2) Please rank the following three aspects in order of how important they were in making the above decision (1 = most important, 3= least important)

Content
Grammar
Organization

3) Could he/she pass an introductory course in your field?

Yes
No

Appendix 3 – Scoring Profile for ESL Instructors

Rater's Initials: _____

Scoring Profile (ESL Instructors)

Instructions: Please fill out the form after reading the attached paragraph. On some of the attached sheets, there is more than one paragraph. Please answer the following questions for the highlighted paragraph only.

1) Do you think this writer has the English language skills to be successful at college-level work?

Yes, definitely
Possibly
Possibly not
Definitely not

2) Please rank the following three aspects in order of how important they were in making the above decision (1 = most important, 3= least important)

Content
Grammar
Organization

Appendix 4 – General Prompt

Write an essay on the following topic:

Some people think that they can learn better by themselves than with a teacher. Others think it is always better to have a teacher. Which do you prefer? Use specific reasons to develop your essay.

You have 30 minutes to complete this task. Your essay will be scored on organization, content, and language.

Essay topic reprinted by permission of Educational Testing Service, the copyright owner.