AUTHOR	Carlson, Sybil B.
TITLE	Relationships of Reasoning and Writing Skills to GRE Analytical Ability Scores. GRE Board Professional Report No. 84-23P.
INSTITUTION	Educational Testing Service, Princeton, N.J.
SPONS AGENCY	Graduate Record Examinations Board, Princeton, N.J.
REPORT NO	ETS-RR-88-13
PUB DATE	Apr 88
NOTE	79p.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC04 Plus Postage.
DESCRIPTORS	Arabic; Chinese; College Entrance Examinations; *College Students; Ethnic Groups; Graduate Study; Higher Education; *Scoring; Spanish; Test Items; *Thinking Skills; *Verbal Ability; Verbal Tests; *Writing Skills
IDENTIFIERS	*Analytical Tests; *Graduate Record Examinations; Language Minorities

ABSTRACT

The reasoning skills tapped by the analytical measure of the Graduate Record Examinations were studied by examining how performance on its constituent type items relate to alternative criteria. Another objective was to ascertain the extent to which additional information on examinees' analytical skills might be obtained from further analyses of their writing performance. The data base consisted of 406 writing samples prepared by 203 examinees (89 native speakers of English and 6 native speakers of Arabic, 73 of Chinese, and 35 of Spanish). Scoring methods were developed that focused on the reasoning skills reflected in the papers. Three scoring methods did not appear to add information beyond that obtained from the analytical reasoning and verbal sections of the GRE General Test, but the scheme developed by P. Moss yielded scores that were relatively independent of these sections of the GRE, possibly tapp ng verbal reasoning skills not assessed by the GRE General Test. Writer's Workbench computerized text analysis suggested that the different writing tasks elicited different kinds of performance, and that the performance of students of different native language groups may vary in complex ways. Four appendixes provide topic and scoring information. (Contains 13 tables and 35 references.) (SLD)

GRE®

GRADUATE RECORD EXAMINATIONS

RELATIONSHIPS OF REASONING AND WRITING SKILLS TO GRE ANALYTICAL ABILITY SCORES

Sybil B. Carlson

GRE Board Professional Report No. 84-23P
ETS Research Report 88-13

April 1988

This report presents the findings of a research project funded by and carried out under the auspices of the Graduate Record Examinations Board.

ETS

EDUCATIONAL TESTING SERVICE, PRINCETON, NJ

Relationships of Reasoning and Writing Skills to GRE Analytical Ability Scores

Sybil B. Carlson

GRE Board Report No. 84-23P

April 1988

Educational Testing Service, Princeton N.J. 08541

ABSTRACT

The major objective of the study was to gain more information about the reasoning skills tapped by the GRE analytical measure by examining how performance on its constituent item types relate to alternative criteria. A second objective was to ascertain the extent to which additional information on examinees' analytical skills might be obtained from further analyses of their writing performance. The data base for this study consisted of 406 writing samples prepared by 203 students who had recently taken the GRE General Test for admission to institutions of higher education in the United States. The bulk of these data were collected for research funded by the GRE Board and TOEFL Policy Council, in which the writing samples were scored to reflect writing skills; these scores were related to scores on the GRE General Test and the TOEFL, as well as other measures. In order to supplement the sample of native speakers, additional writing samples were collected from 77 native speakers of English and 3 native speakers of Arabic who had recently taken the GRE General Test and who were in their first year of graduate education in the United States. The final sample included subsamples of nonnative speakers of English (6 Arabic, 73 Chinese, 35 Spanish) and the subsample of 89 native speakers of English.

The objectives of this study were accomplished by developing several scoring methods that focused on the reasoning skills that are reflected in these papers. These scores, in addition to the scores for writing skills, were related to item type subscores derived from the verbal and analytical reasoning sections of the GRE General Test in order to determine if these item types relate differently to judgments of examinees' thinking and writing skills.

Three scoring methods did not appear to provide additional information beyond what is obtained from the analytical reasoning and verbal sections of the GRE General Test. The Moss scheme, however, yielded scores that were relatively independent of these sections of the GRE. It is possible that these scores tapped verbal reasoning skills that are not assessed by the GRE General Test, but further research is needed to determine whether they represent important developed abilities. Writer's Workbench computerized text analyses suggested that the different writing tasks elicited different kinds of writing performance, and that the writing performance of students representing different native language groups may vary in complex ways in response to these tasks.

ACKNOWLEDGMENTS

This project was the collaborative endeavor of ETS research psychologists and external collaborators. I am very grateful to the brilliance and enthusiasm of the external collaborators: Lawrence T. Frase,, Pamela Moss, Alan C. Purves, Joy Reid, and Anna Soter as well as their professional colleagues who participated in the design and scoring of the different scoring schemes. I also appreciate the data collection efforts made at their institutions by my colleagues: Paul Angelis, Kristina Billmeyer, George Gadda, and Joy Reid. Finally, several members of the ETS professional staff made valuable contributions at various stages of the project: Robert Boldt, Mary Enright, Walter Emmerich, Donald Powers, James Rosso, and Janet Waanders.

Table of Contents

I. STUDY RATIONALE AND DESIGN

The major objective of the study was to gain more information about the reasoning skills being tapped by the analytical measure of the Graduate Record Examinations (GRE) General Test by examining how performance on its constituent item types relate to alternative criteria. A second objective was to ascertain the extent to which additional information on examinees' analytical reasoning skills might be obtained by more detailed analyses of their writing performance.

At its April 1984 meeting, the GRE Board decided to remove the experimental status of the GRE analytical ability measure. In making this decision, the Board anticipated further research and development that would lead to continuing refinements in the analytical measure. The Board also recognized, however, the need to avoid any sudden or radical changes in the measure in order to maintain, if possible, the continuity of the analytical scale over time. This suggests that the two currently used analytical item types, which have thus far proven to be adequate from most psychometric standpoints, will continue to play some role in future editions of the analytical measure.

As has been pointed out, however (Khoury, 1984), there are lingering questions about what the analytical ability section actually measures; specifically, the two analytical item types that make up the current test do not seem to measure a single trait. Instead, one analytical item type (logical reasoning) is more highly related to GRE verbal item types than to the other analytical item type (analytical reasoning). The analytical reasoning type, in turn, is more highly related to quantitative item types than to the logical reasoning type. This situation is explained at least partly by the fact that the two other sections of the GRE General Test also measure reasoning, either with verbal or with quantitative material, but do not attempt to focus explicitly on "pure" reasoning skills. Because of this, one traditional approach to construct validation, the examination of intercorrelations among test items, provides little justification for interpreting the current test as a measure of analytical ability. However, as the new *Joint Technical Standards for Educational and Psychological Testing* (1985) suggest:

> Test validation is the process of *accumulating* evidence to support inferences . . . and there are many ways of accumulating evidence to support any particular inference . . . (p. 1-1)

One such alternative is to examine relationships of a test to other measures purporting to measure the same construct, and to still other measures of different constructs.

Objectives of the Study

The major objective of the study was to provide further information on the construct validity of the GRE analytical measure, and of its constituent item types, by examining their relationships to alternative criteria. Specifically, the study explored whether the pattern of correlations among the two GRE analytical item types, the GRE verbal item types, and several scores derived from writing samples might provide evidence that analytical items reflect analytical or reasoning skills in ways that are not reflected by other GRE item types. In general, we would expect to observe patterns such that, when related to "reasoning" and "writing" scores derived from the writing samples, analytical item types would correlate relatively better with thinking scores than with writing scores, when compared with correlations for verbal item types. A secondary objective was to ascertain the extent to which additional information on examinees' analytical skills might be obtained from further analyses of their writing performance.

Background

When an individual solves a problem, his or her ideation is mediated by, and couched in, verbal forms. When an individual writes a composition, the process of expressing his or her thoughts is essentially a problem-solving activity that is reflected by the organization, level of conceptualization, and selection of terminology that are presented in the written product. Furthermore, the quality of the written composition is evaluated by others in terms of the degree to which it effectively communicates ideas. This judgment of competence may be influenced both by the individual's mastery of the conventions of writing and by his or her developed reasoning skills. In the context of standardized testing, these problem-solving and writing skills traditionally are assessed by items cast in a multiple-choice format that indirectly assess these skills through the recognition of a correct response among a fixed set of possible responses. In contrast, an open-ended test item attempts to assess these skills more directly by requiring the production of responses.

A series of research projects previously conducted for the Graduate Record Examinations Board investigated the measurement properties of test items designed to assess problem-solving skills in multiple-choice and open-ended formats (Ward, Frederiksen, & Carlson, 1980). A construct validation study suggested that the open-ended Formulating Hypotheses (FH) items tapped cognitive skills that are not revealed in response to multiple-choice items--specifically the ability to produce ideas. The FH item required candidates to list as many ideas, or hypotheses, that might provide an explanation of a problem, as they could think of. The items were scored by judges using a categorical system to reflect the quality of responses, from which a number of different scores were derived. Judges were cautioned against being influenced by the writing ability of the examinee, and were asked instead to focus on assigning category numbers that corresponded to the ideas being expressed, however poorly. Given this focus, examinees who expressed good ideas would not be penalized unduly for inept writing skills. In addition, the examinee was instructed to list a number of discrete ideas, rather than to combine the ideas into a developed composition; hence the examinee's ability to organize the ideas and to communicate them effectively through writing were not evaluated. This previous GRE Board-sponsored research demonstrated (1) the feasibility of obtaining multiple scores from a single task, and (2) the likelihood that open-ended items may tap reasoning skills that are not captured by multiple-choice items.

More recently, the GRE Board and TOEFL Policy Council have supported a study of the relationships of scores on sections of the GRE General Test and the TOEFL to a variety of scores on direct measures of writing, or writing samples (Carlson, Bridgeman, Camp, & Waanders, 1985). The scoring of these writing samples focused on the evaluation of writing ability from the perspective of academic competency in written English. The data collected for this study consist of approximately 2,500 papers written in response to four different topics by 638 candidates for admission to higher education institutions in the United States. The writing samples were generated by undergraduate and graduate candidates in countries that represent three predominant language groups: Chinese, Spanish, and Arabic; writing samples also were written by a sample of English-speaking candidates seeking admission to graduate schools in the United States. These papers were scored to assess writing ability in several ways: (1) holistically (both for one overall impression score and two scores for discourse/sentence skills) by English and English-as-a-second-language (ESL) instructors/readers, (2) by graduate faculty in two diverse disciplines, and (3) by the Bell Labs' Writer's Workbench software system to obtain a large number of numerical indices (e.g., length of essay, number of spelling errors, percentage of vague words).

The writing stimuli, or assignments, to which the subjects for this study responded were carefully designed to elicit forms of writing that are valued in institutions of higher education in the United States and Canada, as determined in an earlier survey (Bridgeman & Carlson, 1983) of academic writing tasks. In that study, faculty members from a representative sample of departments and universities rated several types of topics that might be used to assess the writing ability of entry-level undergraduate and graduate students. Although faculty from the different disciplines did not agree completely on the most appropriate writing stimuli, they indicated a preference for two types of topics that appear to place somewhat different demands on the writer. These topics require writers to (a) "compare and contrast plus take a position" and (b) "describe and interpret a chart or graph." The faculty members indicated that they had selected these topics because they would provide, in addition to an assessment of writing skills, some evidence for the thinking skills of the candidates. The compare/contrast and chart/graph topics clearly demand the application of certain thinking skills that structure the organization and content of compositions written in response to the topics. In the compare/contrast mode, the writer must organize ideas related to two aspects of a situation and present them in the form of a persuasive argument. In the chart/graph mode, the writer must comprehend and interpret a visual/verbal stimulus and discuss the information clearly. Furthermore, these thinking and writing skills demanded by the two topics reflect academic competencies that are required of graduate students.

Some of the pertinent results and conclusions obtained from the GRE/TOEFL project were the following:

- o When the readers were being trained in holistic scoring procedures, their comments indicated that their scores were heavily influenced by the quality of overall organization of the papers. They noted that this feature of organization was dependent on paragraph- and sentence-level cohesion, rather than on mechanical characteristics; in turn, cohesion was influenced by the quality of thinking that entered into the composition (e.g., linear reasoning, poor deductive logic). This observation is supported by recent research (Breland & Jones, 1982) in which a special scoring system was used to code the characteristics of writing. The results indicated that characteristics such as organization, transitions, use of supporting evidence, and the originality of ideas had more impact on holistic scores than did such syntactical features as subject-verb agreement, punctuation, and pronoun usage.

- o Some of the reader comments also acknowledged their appreciation of cross-cultural differences for which papers should not be penalized--different approaches to the logical organization of the ideas that reflected an acceptable practice in another, non-Western culture.

- The estimates of reliability of holistic scores assigned to the same paper by different readers were consistently high (.80-.85),* indicating they were in close agreement regarding the criteria that influence general impressions of writing skills. When holistic scores were correlated within and across topic types, the correlations were no higher within topic types than across topic types. This finding supports the view that, at least for these topics, there were not systematic differences in the ranking of student scores for papers written in response to different topic types. This finding provided justification for limiting, for cost efficiency, the scoring of papers for the proposed study to papers written in response to two topics, one of each topic type.

- These high estimates of reader reliabilities assigned by different readers indicate that the readers were able to reach considerable agreement on the relative quality of the papers they were judging. However, this evidence does not indicate whether readers are evaluating the same features of writing or whether they are attending to different features when making decisions to assign specific scores to writing samples that require different approaches to the task (e.g., compare/contrast vs. chart/graph). In fact, the readers did report that the overall writing performance of candidates on the chart/graph topics was not as high in quality as performance on the compare/contrast topics. If verified by alternative scoring methods, this performance differential might be explained by the differences in the cognitive demands of the two tasks.

- The means of the writing sample scores reflected consistent level differences for the three language groups for whom English is not the primary language. For every writing sample score, the means were lowest for the Arabic sample, in the middle for the Chinese sample and highest for the Spanish sample. Since a portion of the data base for the present study consists of writing sample and GRE scores for nonnative speakers of English, primarily Chinese and Spanish, a more detailed analysis of these writing samples might suggest some explanations for these level differences.

*Spearman-Brown correction providing an estimate of the reliability of the scores based on summing the judgments of two raters.

- A principal axes factor analysis with varimax rotations of holistic scores and TOEFL section scores resulted in a two-factor solution. The two factors appear to be method factors, one consisting of scores on the three sections of the TOEFL, and the other, of holistic scores on papers written in response to the four topics. One interpretation of the two factors suggests that performance on measures of English language proficiency becomes more differentiated when English proficiency measures require a candidate to respond by applying different cognitive processes--recognition vs. production. This result provides further confirmation that item formats in which examinees generate reponses yield information about performance not yielded by multiple-choice formats.

- Representative subsamples of papers written on the one topic of each type were mailed to four graduate-level social science professors and four engineering professors. They were instructed to rate each set of papers on a 1-6 scale from the perspective of academic writing competence expected of beginning graduate students. The professors' ratings were highly correlated with each other--the mean social science ratings correlated .92 with the mean engineering ratings for each of the two topics. When compared with the holistic scores assigned during the regular scoring session for the compare/contrast topic, the mean social science judgment correlated .86 with the holistic scores, and the mean engineering judgment, .92. For the chart/graph topic, the correlations were .83 and .82, respectively. This outcome supports the assumption that general agreement exists, even when not formally identified and verbalized, concerning standards for academic writing competence. It also provides some justification, for the present study, for mailing the writing samples to professionals who will score them for reasoning skills at their institutions, rather than bringing together and training a group of readers at ETS.

- When the holistic writing sample scores, averaged over four topics, were related to scores on item types within the sections of the GRE General Test, the observed pattern of correlations was consistent with the relationships reported in other GRE studies (Table 1). Specifically, the analytical reasoning and logical reasoning scores were not highly correlated (.23), and the analytical reasoning items were more highly correlated with the quantitative items (.46, .35, .50) than were the logical reasoning items (-.09,

-.18, .02). On the other hand, the logical reasoning items were more highly correlated with the verbal items (.65, .50, .67) than were the analytical reasoning items (.15, .17, .24). The holistic scores were more highly correlated (.64) with the logical reasoning items than with the analytical reasoning items (.23). This result indicates that the holistic scores, as expected, reflect verbal ability, and suggests that it would be interesting to determine whether a scoring system that focuses on analytical reasoning skills evidenced in the writing samples would yield higher correlations with the analytical reasoning items on the GRE General Test.

The results of the study demonstrated that the writing performance of nonnative speakers of English can be evaluated reliably, and that direct measures of writing performance, although moderately correlated with multiple-choice measures, contribute additional information regarding the English proficiency of foreign candidates.

The Current Study

In the previous study, one aspect that clearly distinguished levels of writing competence, and that was apparent to the investigators and scorers in reading the writing samples, was the quality of the reasoning expressed in the papers. In the postsecondary contexts in which students write papers to communicate their ideas within a discipline, educators evaluate observable features of the written text in order to make judgments about the quality of student thought. This situation poses a practical measurement concern. Because academic grades are assigned on the basis of such judgments, and because these judgments reflect implicit cultural norms, this research focuses on making explicit the criteria that influence the judgments. Although this study is exploratory, based on somewhat restricted data and sample sizes, it is intended to stimulate further investigation. As described in the next section of this paper, we attempted to develop a variety of scoring schemes that reflected the perceptions of different communities of readers regarding the reasoning skills that could be observed, reliably and with validity, in the written discourse of native and nonnative speakers of English. The scores yielded by these schemes were analyzed in relation to one another as well as in relation to scores on verbal and analytical item types of the GRE General Test. Ultimately, one or more elements of such schemes might contribute to our understanding of the differential performance of writers from different cultures and, possibly, toward effective approaches to feedback in instruction when writing skills are exercised.

In using the term "reasoning skills," we do not assume that the writing sample data will provide information about the thinking processes applied by the writers, or that we can develop scores that describe sophisticated constructs such as analytic reasoning. We recognize that the processes involved in thinking and writing are investigated more legitimately through protocol analysis, or "think-aloud" studies. In addition, we understand that the quality of thinking can only be inferred indirectly when the product of that thinking is in written form, and that, in fact, poorly developed writing skills can mask an individual's ability to express thoughts verbally.

We did not approach this task by developing a hypothetical model of the complex factors that might contribute to writing that might be judged "different," but equivalently competent within a crosscultural perspective. Instead, this research is exploratory, intended to obtain more information that may lead to hypotheses to be tested and, eventually, to possible models to be tested that might be useful to the evaluation and instruction of writing. An accumulated body of research data might yield systematic patterns of relationships among the variables we have created and labeled, therefore contributing to the inferences made about the validity of the model(s).

Data Collection

The data for this study consisted of 406 writing samples prepared by 203 students who had recently taken the GRE General Test for admission to institutions of higher education in the United States. The students in the sample consisted primarily of nonnative speakers of English whose native language is Arabic (6), Chinese (73), Spanish (35), and a sample of native speakers of English (89).* The bulk of these data were collected for the previous GRE/TOEFL project (Carlson et al., 1985), in which papers written by each of 132 students on four topics were scored to reflect writi g skills. To supplement the original sample of native speakers, additional writing samples were collected from 77 speakers of English and 3 students whose native language was Arabic.

These data were collected by a campus faculty member at each of four sites: Colorado State University, the University of California at Los Angeles, the University of Pennsylvania, and Southern Illinois University. Each student was paid $30 to write essays in response to the two topics with a 30-minute time limit

*Henceforth, native speakers of the different languages will be referred to as the Arab, Chinese, English, and Spanish samples for speakers of those respective languages.

per topic. Although we attempted to obtain more writing samples prepared by Arabic speakers, our campus representatives were able to recruit only three. Altogether, data for 89 additional students were collected, but GRE General Test scores could not be found for 8 students and the papers written by one student were unscorable, resulting in the complement of 80 papers to the data obtained in the original study.

Section II of this report discusses the development and application of the reasoning scoring schemes. Section III describes the results, and Section IV, the conclusions.

II. DEVELOPMENT OF THE SCORING SCHEMES

On the basis of the findings of the writing survey study, two types of topics were selected to serve as complements to one another, particularly since the demands of each task would be expected to elicit evidence of different writing skills that are required in postsecondary academic writing. For the GRE/TOEFL writing assessment study, writing samples written in response to four topics, two of each topic type (compare/contrast and chart/graph) were collected.

The results of the GRE/TOEFL study indicated that correlations among holistic scores were as high across topic types as within topic types. This result suggests that (1) the different topics did not elicit qualitatively different writing performance and/or (2) the readers maintained a comparable scale for evaluating the writing samples, despite performance fluctuations from topic to topic. These results should not be interpreted as evidence that papers written in response to any topic or topic type would yield equivalent scoring agreement. The correlations reflect the relative, rather than absolute, values of student scores (e.g., a student with a high score on one topic received a high score on the other topic). Since a holistic score summarizes an overall, general impression, different features of writing may have contributed to the same score for different individuals as well as for different tasks. Although readers could learn to internalize criteria for reliably evaluating writing samples within a specific context, it is likely that performance varies both in degree and kind in response to different task demands. Because the single scores are likely to reflect different combinations of features, it is possible that important differences might be revealed if multiple scores were used.

Because the data analyses indicated that the holistic writing scores within and across topics were highly correlated, the writing samples obtained for the two contrasting topic types, Space and Farming (Appendix A), were used for further analyses in the current study. The Space prompt, a compare/contrast topic type, asked students to write an essay comparing the advantages and disadvantages of the exploration of outer space and to take a position. The Farming topic, a chart/graph topic type, presented three graphs (number of farms, size of farm population, average farm sizes) depicting changes in farming in the United States from 1940 to 1980. The writers were asked to write a report that interpreted and interrelated the graphs and explained the conclusions reached from the information in the graphs, using the graphs to support the conclusions. All writing samples were collected under controlled examination conditions with a 30-minute time limit per task. For the nonnative speakers in the original sample, the papers were written during the afternoon of the same day the candidates took the TOEFL.

At the outset, we recognized that the papers, obtained for somewhat different purposes, had limitations but would serve the purposes of this exploratory study. Ideally, the demands of the writing tasks might have been different and more explicit. For example, the instructions to the student writers did not indicate that reasoning skills were to be evaluated. Also, the task demands might have been structured differently to elicit specific reasoning skills. Despite these drawbacks, the project staff and collaborators agreed that some reasoning skills could be identified since they contributed to the organization of the ideas developed in the papers. Another limitation was posed by the relatively small sample sizes, particularly if broken down by major field and native language groups. As a result, the data collected from the small subgroup of Arabic speakers were only included in the total-group analyses, and the score data were not analyzed by major fields.

The Scoring Schemes

In order to provide the kinds of information that are needed to fulfill the purpose(s) of a writing assessment program, evaluators have developed a range of different scoring schemes. Holistic scoring in its various forms recognizes that the singular factors that constitute a piece of writing are elements that work together to make a total impression on the reader. Holistic scores, though, do not provide sufficient information for diagnostic purposes or for research that will lead us to a better understanding of the interplay of the numerous variables that influence and contribute to a definition of the writing skills being evaluated. As a result, many different scoring schemes have been devised in order to obtain multiple scores. Typically, these schemes tend to oversimplify and mechanize the evaluation task by adopting or inventing a proliferation of labels that reflect a particular reader perspective. These labels do not necessarily reflect the perceptions of another community of readers who do not share the same perspective. The readers who share a common perspective may produce highly reliable judgments in evaluating writing samples, but the validity of the scores assigned to these labels is questionable unless readers who represent a different audience can agree with these judgments as well. Making a match between the expectations of the reader and the writer involves yet another complex interaction, since readers and writers have acquired their idiosyncratic approaches to defining competent performance largely through educational experience within their particular cultures or subcultures.

For the current study, several experts in writing assessment and linguistics, with experience in research and instruction in reasoning/writing skills and in English as a second language

(ESL), were convened for two days in May 1985 to discuss the objectives of the research and to formulate strategies for developing scoring schemes that would emphasize the reasoning skills that might be observed in the writing samples. Three of these schemes are named after their primary developers. The following scoring schemes were applied to the writing samples: the Purves/Soter scheme, the Moss scheme, the Reid holistic reasoning scheme, the holistic writing scheme, and the UNIX Writer's Workbench (WWB) system* supplemented by Lawrence Frase's computer analyses.

The two holistic schemes represent approaches to evaluation for which the criteria being applied are least explicit. Although the readers used benchmark papers to maintain comparable standards, the benchmarks provide examples in which the integration of potentially different and discriminable features contribute to one numerical score. The Purves/Soter and Moss schemes attempt to evaluate several characteristics of the reasoning process demonstrated in written prose that might be identified by human judges. At the extreme of the range of scoring specificity, the WWB programs yield a considerable number of explicit features of the texts.

The Purves/Soter Scheme

Alan Purves and Anna Soter, then at the University of Illinois in Champaign-Urbana, conducted preliminary analyses of 20 representative writing samples for the Space and Farming writing topics. They adapted, modified, and extended Carroll's (1960) descriptive vectors of prose style, from which they derived 14 unipolar scales that were applied to the papers (Appendix B). A factor analysis yielded three factors that reflected content/thinking, organization, and style/tone as identifiable features of the essays. This work culminated in the development of the scoring scheme (Purves, 1985) used for both the Space and Farming papers in this study. Briefly, the scheme identifies the following features:

1. Content/Thinking
   a. Adequacy of information presented
   b. Richness of additional information
   c. Relationships drawn
   d. Inferences made
   e. Synthesis
   f. Evaluation
   g. Consideration of alternatives

*Trademark of AT & T

2. Organization
   a. Framing
   b. Grouping
   c. Unity

3. Style/Tone
   a. Objectivity
   b. Tentativeness
   c. Metalanguage

The specific aspects of each of the three major dimensions were evaluated separately, using a 1-5 scale with "1" as the lowest rating. These features are further defined in greater detail; Anna Soter developed an even more extensive scoring guide, which also contained examples from the papers (Soter, 1985).

The Space and Farming writing samples were rated by four experts at the University of Illinois with knowledge and experience in the evaluation of English and ESL writing. They were trained in two sessions in which the criteria were discussed in relation to 10 samples per topic. During their evaluation of the full set of papers, the readers referred to a chart of categories and criteria. All four readers rated all papers for the Space topic. For the Farming topic, two different groups of two readers rated the odd-numbered papers and two readers rated the even-numbered papers.

Estimates of reader reliability of scores assigned to the separate categories within each dimension were quite low. These separate categories, however, are assumed to represent elements that contribute to and define each of the three major dimensions. Thus, the separate subscores assigned by each reader were summed within each dimension, and reader reliability coefficients (alpha) were calculated for the three dimensions for each topic. The reader reliability estimates for the Space topic were considerably higher than for the Farming topic: content/thinking, .88; organization, .85; style/tone, .85. For the Farming topic, the reader reliabilities were as follows: content/thinking, .66; organization, .63; style/tone, .66. These reliability coefficients for the Space topic fall within a range that is acceptable in the evaluation of open-ended responses by human judges. For the Farming topic, however, the readers appeared to experience greater difficulty in making reliable judgments about the features of these papers. Because the reader reliability estimates were not abysmally low, but only lower than what is considered highly acceptable, the Farming scores were still included for the purposes of this study. These reliabilities would not be acceptable in any testing or other decision-making situation, but this topic still might provide some useful information in this exploration, as long as interpretations regarding the Farming data take into account the lower reliabilities of scores.

For the final analyses, the scores assigned by each reader, summed over each of the three dimensions for each of the two topics, were averaged to obtain one score per student for each dimension for each topic (a total of six scores per student).

The Moss Scheme

Pamela Moss, of the School of Education of the University of Pittsburgh and director of evaluation for the Critical Thinking Project of the Pittsburgh Public Schools, developed a scheme that builds on the work of Stephan Toulmin and his colleagues (1984). Toulmin uses the word "argument" in a broad sense to mean any claim and the reasons that support or justify it. He has described the features of all arguments as follows: the claim is the statement of one's ideas; grounds provide the foundation of evidence for the claim; warrants provide the rules, laws, formulas, or principles on which the grounds are based; modal qualities indicate the degree of certainty with which the argument is made; and possible rebuttals deal with circumstances where the argument or claim may break down. For the writing tasks in this research, advantages and disadvantages of space exploration, the statement of the writer's position, and the conclusions drawn from these elements were all considered to be claims. For Toulmin, these elements are generic to all arguments, but manifest themselves in different ways in different disciplines (Moss, personal communication, 1986).

With Toulmin's scheme as an approach to organizing data regarding reasoning skills, Moss read through sample papers in order to distill the ways in which these elements might be manifested in the discourse produced in response to the two topics. The scoring scheme she designed (Appendix C) is intended to identify the elements relevant to thinking or reasoning and to ignore those aspects of the essays--focus, organization, irrelevancies, redundancies--that usually are taken into account when evaluating writing and may be differentially valued in different cultures.

Related but specific scoring schemes were developed for the two topics since reasoning skills were evidenced in different ways in response to the specific demands of each topic type. Also, the labels for the elements of the scheme were couched in more common terms, eliminating the need for readers to be familiar with the Toulmin approach. Essentially, readers were asked to list all claims made in the essays and then to evaluate each claim independently. Emphasis was placed on ideas in context rather than on the sequence of ideas in order to stress reasoning skills and not the writing skill of organization. Instead of evaluating the characteristics of the papers, readers indicated the presence (or absence) of the particular features, which resulted in scores that reflected frequencies of occurrence. Five readers, all experts in the instruction and evaluation of English and ESL writing skills, spent five days scoring the entire set of papers using the Moss scheme. Approximately four hours per topic were required for the training of the readers by the project director; the readings required slightly more than seven hours per topic. The training involved working with sets of sample papers in order for each reader to arrive at internalized definitions of the reasoning skills in the scheme as they are evidenced in written discourse (e.g., to agree on the kinds of performance that indicate a "reasonable inference").

The tallies on the score sheets were compressed to reflect the major elements of reasoning skills the scheme was designed to identify. Four scores were derived for the Space (compare/contrast) topic:

1. Claims
2. Support (justifications)--evidence and explanation
3. Qualifications/rebuttals
4. Integration--claims as grounds for subsequent claims

For the Farming (chart/graph) topic, six scores were derived, four that were parallel to those for the Space topic, and two (1 and 2) that are specific to the topic type:

1. Graph reading skills
2. Deduction skills
3. Claims
4. Justifications
5. Qualifications/rebuttals
6. Integration

The score derivations, excepting the two unique Farming scores, involved dividing the tallies for each score by the tallies reflecting the number of ideas to eliminate fluency in evaluating the quality of reasoning skills.

Estimates of reader reliability (Kuder-Richardson) for the Moss scheme were higher for the Farming topic than for the Space topic, the reverse of the trend for the reliability estimates for the Purves scheme. For the Farming topic, the reliability coefficients were as follows: graph reading, .81; deductions, .62; claims, .71; justifications, .55; qualifications/rebuttals, .45; and integration, .30. It is interesting to note that the indicators of graph reading skills were easier for readers to identify (.81 reliability). As the elements of reasoning skills became more elusive and less concrete (e.g., qualifications/

rebuttals and integrations), they were identified less reliably. Another reason for the lower reliability of these elements is that very few writing samples evidenced these higher-order reasoning skills; slight disagreements among readers regarding the small number of papers demonstrating a particular characteristic are likely to dramatically influence estimates of agreement. Given these low reliabilities, however, the integration score was eliminated from the Moss scheme analyses for the Farming topic.

Estimates of reader reliability for the Space topic were as follows: claims, .59; support, .61; integration, -.02; and qualifications, .51. The extremely low coefficients for integration agair can be explained by the very small number of writers that evidenced this aspect of reasoning. The score for integration consequently was drcpped from the analyses.

For the final analyses, the scores assigned by the two readers were averaged to obtain one score for each element of the Moss scoring scheme for each writer on each topic. As with the Purves/Soter scheme, scores with only moderate reader reliabilities were retained, but the reliabilities need to be taken into account when interpreting the data.

The Reid Holistic Scoring Scheme

Joy Reid, associate professor in the Intensive English Program of the Department of English at Colorado State University, has considerable experience in ESL writing instruction and research, and in evaluating the compositions of ESL writers, both as a reader and using the Writer's Workbench. Working with sample papers and with ETS writing assessment expert Janet Waanders, she developed a broad rubric to guide readers who evaluated the papers holistically, with an emphasis on reasoning rather than on writing skills (Appendix D). The same readers who applied the Moss scheme were trained to assign the general-impression scores by using benchmark papers that reflected the range of the 1-6 score scale (with "1" as the lowest score). A holistic reasoning score was assigned by each of two readers per paper before the readers applied the Moss scheme to the paper. Although the overall rating of the papers may have partially influenced the readers' Moss evaluations, the readers felt that they perceived an overall rating on the first reading, regardless, and that the use of the Moss scheme required a different kind of decision making in that the treatment is more analytical and does not involve judgments about degrees of quality or sequence.

Estimates of reader reliability (coefficient alpha) for both topics for the holistic reasoning scores were highly acceptable: .88 for the Space topic and .86 foι the Farming topic. For the final analyses, the scores were averaged across the two readers, resulting in one holistic reasoning score per writer per topic.

The Holistic Scoring of Writing Skills

A major portion of the writing samples had been holistically scored for the evaluation of writing skills for the TOEFL/GRE study. A full description of the scoring appears in the report of that research (Carlson et al., 1985). This method was duplicated in the scoring of the 80 additional papers per topic that were collected to supplement the writing samples for this study. These papers were read in less than one day by two readers, one an ESL writing expert, the other an English writing expert, who were trained by a chief reader. All three readers had participated as readers in the TOEFL/GRE study, and were trained with the same benchmark papers used previously in order to maintain the same standards. Each paper was read by two readers, who required less than one half day to evaluate the Space and Farming topics separately.

Estimates of reader reliability (coefficient alpha) for the holistic writing scores for both topics were acceptably high: .89 for the Space topic and .92 for the Farming topic. The scores were averaged over the two readers for the final analyses, yielding one holistic writing score per writer per topic.

Differential Reader Reliability Estimates for Language Groups

Of the four scoring schemes that required human judgments, the estimates of reader reliability differed for separate language groups and from the total sample estimates, without apparent systematicity, with the exception of estimates for the holistic reasoning scores. For the holistic reasoning scores, estimates of reader reliability were approximately equivalent, except that the reliability of the scores for the Chinese students for the Space topic was somewhat higher than for the other language groups. The reader reliability estimates for the holistic writing scores for all language groups but the English speakers fell within the same range. The reliability estimates for the English group undoubtedly were low because of restriction of range--their holistic writing scores were all very high. For the Purves/Soter scheme, the reader reliability estimates were lower than those for the other language groups for the content/thinking and organization scores for the Space topic, and for organization scores for the Farming topic. Reliability estimates for both the English and Spanish groups were somewhat lower than those for the Chinese group for the content/thinking and style/tone scores for the Space topic. The reliability estimates for the Moss scheme were similarly and unsystematically variable for the different language groups. For the claims and qualifications scores for the Space topic, the reader reliability estimates for the Chinese were higher than for the other language groups, but lower for the support score. For the deductions, claims, and justifications scores for the Farming topic, the estimates were higher for the Spanish group.

The differential reliability estimates observed for these data should not be viewed as generalizable trends. The exploratory nature of the scoring schemes, the relatively small language group samples, and the task demands all probably contributed in undetermined ways to these differences. They are noted, however, because subsequent research should investigate whether reader reliability estimates differ by language group, and whether they do so systematically. Perhaps readers experience greater ease or difficulty in identifying and evaluating certain characteristics of written prose that may be influenced by the different approaches taken by writers from different cultures.

Writer's Workbench Textual Analyses

The UNIX Writer's Workbench system, a computerized text analysis tool, was used in the TOEFL/GRE study to obtain detailed information about the features of the texts for a representative subsample of the writing samples. The data suggested that certain characteristics of writing that are attended to by a human reader are related to, and therefore are likely to influence, the evaluation (in this case, holistic evaluation) of a piece of writing. It is possible that readers respond unconsciously to particular features of writing that affect their judgments. Thus, we realized that the WWB programs could contribute to our understanding of the implicit criteria that are applied in holistic scoring. Furthermore, we recognized that multiple scores assigned by human judges might yield additional insights about these criteria. Therefore, in the current study, we applied the Workbench analyses to our newly designed scoring schemes.

The Writer's Workbench software programs at Colorado State University (CSU) were used to analyze all writing samples. A writing instructor was employed to type all papers for computer input. Her work was edited for fidelity to the actual papers, and the WWB output was obtained. In addition, Lawrence Frase, who contributed to the development of the Workbench programs at AT&T Bell Labs, provided advice and conducted other exploratory analyses.

The WWB programs incorporate expert knowledge from rhetoric and psychology, offer a variety of approaches to assessing writing skills, and can be adjusted to the standards for a particular writing population (Cherry et al., 1983; Kiefer & Smith, 1984; Smith & Kiefer, 1982).

The CSU programs have been refined over several years to reflect the standards for academic writing at the college level (Reid & Findlay, 1986). The WWB output yields a considerable amount of data with which to describe the syntactic and semantic features of written prose in numerical or quantifiable forms. Many of these features were highly interdependent (e.g., number of words), infrequently observed in these data (e.g., compound-complex sentences), or unnecessarily specific. For our purposes, we focused on certain features that might provide information relevant to reasoning/writing skills. The number of features to be used in the analyses was further reduced on the basis of the intercorrelations of the features: features that were highly correlated, hence very interdependent, were reduced to one representative feature, and some features were eliminated if they were identified infrequently in this set of papers (see list in Table 1). For example, the percentage of shorter sentences was eliminated in favor of the percentage of longer sentences and average sentence length. Many of these features are self-explanatory and serve as generally agreed-upon indicators of good and less effective writing skills; more complete definitions appear in Reid and Findlay (1986) and the other WWB references. Those features most relevant to the data analyses in this paper are described in the section reporting the results.

The efficiency and reliability of a computerized text analysis system, of course, are high in relation to those of human judges. Because the programs lack the human judgment needed to identify certain aspects of English prose, however, the programs are not 100 percent reliable. The parts of speech program, for example, is approximately 90 percent accurate; the identification of passive voice is 93 percent accurate (Cherry et al., 1983; Frase, in press). Also, the program that flags spelling errors is not always accurate. For this study, words that actually were not misspelled but were listed in the output as such were eliminated from the counts.

GRE Part Scores

The GRE General Test is composed of three major sections, verbal, quantitative, and analytical reasoning. Because the emphasis of the study was on verbal and reasoning skills, the quantitative scores were not included in the analyses. Separate part scores for the GRE General Test are not reported to candidates because their contribution to the instrument depends on the combination of the parts. For research purposes, though, we were able to calculate separate scores for the parts of the verbal and analytical reasoning sections. Three parts of the verbal section--sentence completion, discrete verbal, and reading comprehension--assess different aspects of language skills. The two parts of the analytical reasoning section represen two important aspects of the reasoning process (GRE Guide to the Use of the Graduate Examinations Program, 1986-87). The analytical reasoning items assess the ability to understand a structure of arbitrary relationships, to deduce new information from the relationships given, and to assess the conditions used to establish the structure of the relationships. The task demands of the Space topic, in contrast, require the writer to generate a structure of relationships and move beyond the brief structure that is supplied by the prompt. The Farming topic, on the other hand, offers some structure that writers can used to organize the development of their ideas.

The logical reasoning items assess the ability to understand, evaluate, and analyze arguments. Studies validating the analytical reasoning section have indicated that logical reasoning scores tend to be at least moderately or highly correlated with the verbal section scores, whereas the analytical reasoning scores tend to be correlated moderately with the quantitative scores (Wilson, 1985).

III. RESULTS

Performance of the Separate Scoring Schemes

The Purves/Soter Scheme

The scores for the three dimensions for each of the two topics were correlated (Table 3)* to determine the degree to which the dimensions are related. Within one topic type, the correlations among the three dimensions were very high, ranging from .90 to .93 for the Space topic and from .88 to .91 for the Farming topic. Across topics, the correlations among the three dimensions were moderately high, ranging from .72 to .77. Thus, within a topic, scores assigned to three dimensions appear to be highly interdependent; it would be difficult to justify a claim that any one of the dimensions is measuring something that is not being measured by the others. The higher correlations within a topic as opposed to across the two topics suggest that the readers reacted to the specific demands of the different topics when evaluating the papers.

The scores obtained for the three Purves/Soter dimensions also were highly correlated with the holistic writing scores. Correlations of the scores for the three Purves/Soter dimensions per topic with the scores for the Moss scheme yielded relatively low correlations, generally ranging from .22 to .57. The only correlations among these scores that were relatively high were the correlations of the three Purves/Soter scores for Space with the Moss support score for Space (.71 to .72). The high correlations of the Purves/Soter scores with one another and with holistic writing scores suggests that written language ability heavily influences these scores and that reasoning ability has not been isolated from writing ability. The considerably lower correlations with the Moss scores suggest that something other than writing ability has been identified with the Moss scoring scheme.

The scores for the 13 separate elements per topic within the dimensions of the Purves/Soter scheme were factor analyzed as well. The varimax factor analysis resulted in a two-factor solution in which the two factors were defined by the two different writing tasks, Space and Farming. This analysis lent additional support to the unidimensionality of the three dimensions for scoring within one topic, and to the differentiation of the two tasks.

*All correlations reported are at the .01 or .001 level of significance.

Finally, the Purves/Soter scores for the three dimensions for each topic were correlated with the GRE part scores. These correlations yielded the same pattern of relationships, at approximately the same levels, observed for the correlations of holistic writing scores with the GRE part scores (Table 3), reflecting the trend observed with other verbal data and these GRE part scores. The correlations of the holistic writing scores yielded similar patterns and values of relationships. These patterns of relationships underscore the assumption that the Purves/Soter scores are highly related to holistic scores of writing ability and moderately related to verbal ability scores. Although the three dimensions of the Purves/Soter scheme did not appear to be independent, the three scores per topic were retained for final analyses in order to further explore their relationships with Writer's Workbench variables.

The Moss Scheme

Correlations of the separate scores derived from the Moss scheme with one another indicated that the scores were relatively independent of one another (Table 3). The two highest correlations, .55 for support with claims for the Space topic, and .64 for justifications (a form of support) with claims for the Farming topic indicate a necessarily dependent relationship, since support is offered for the claims being made. The correlations of Moss scores across the two topics also were low, another indication of the differential task demands of the topics. Generally, the correlations of the Moss scores with the holistic writing scores for both topics were low to moderate, ranging from .23 to .45, with the exception of the moderate correlations of the Moss support score for Space with the holistic writing score for Space (.65) and the holistic writing score for Farming (.60). Clearly, writing skills play some role in the ability to offer support for a claim.

Correlations of the Moss scores with the GRE General Test part scores resulted in low and moderate correlations. Thus it is possible (a question for further investigation) that the Moss scores reflect verbal skills to some extent, as would be expected, and also tap other abilities or, at least, abilities that are not assessed independently in the GRE analytical reasoning section. An extended discussion of the implications suggested by the relationships of the Moss scores to scores on the GRE verbal and analytical reasoning sections appears at the conclusion of this section.

The Reid Holistic Reasoning Scores

Correlations of the holistic scores for Space and Farming were high within topic (.80 for Space holistic reasoning with Space holistic writing and .82 for Farming holistic reasoning with Farming holistic writing). It is interesting to note that the correlations of holistic writing scores for Space with Farming (.85) and the correlations of holistic reasoning scores for Space with Farming (.74) appear to indicate that holistic writing scores were still more highly related to each other than were the holistic reasoning scores. Because the holistic writing scores and the holistic reasoning scores were relatively highly correlated, the high interdependence of these scores should be taken into account when interpreting the final data analyses.

Finally, the correlations of the holistic reasoning scores further support the assumption that the holistic reasoning and writing scores are essentially interchangeable, since their correlations with the GRE verbal scores reflect nearly identical patterns and levels of relationships.

The Writer's Workbench Analyses

The correlations of the WWB features with one another and with the scores yielded by the scoring schemes reflected some complex relationships beyond the purview of this paper. It is interesting to note that, even though the different Workbench variables would be expected to be highly related, the variables we selected for the analyses did not reflect high correlations with one another for this sample of papers. The WWB thus provides a valuable investigative tool to enable researchers to define the labels they have assigned to features of written discourse. Because the sample size is small, and the number of intercorrelations so numerous, any attempt to generalize from these data would be spurious. In addition, the reliabilities of some of the scores for two of the scoring schemes were not sufficiently high for this kind of analysis.

Varimax factor analyses of the 21 WWB scores for each topic yielded four factors that appear to be interpretable as well as relevant to the performance observed in the papers. The factors reported in Table 4 (eliminating loadings of lower than .30) again support the differences in performance that are elicited by the two topics, but they also suggest that some similar features contribute to the writing skills demonstrated in both topics. Factor 1, for example, depends predominantly on content-related features, although average word length loads on this factor for Farming, and on Factor 2 for Space. Factor 2 seems to reflect sentence variety. Factor 3 again reflects some similarities across the two topics (average word length, average length of content words, percentage of nominalizations, a negative loading for percentage of vague words), but some differences in that additional features are reflected in performance on the Space topic (passive voice, "to be" verbs, number of possible confused homophones and word pairs to check, average sentence length, percentage of nominalizations). Factor 4 further distinguishes between the two topics in several ways: passive voice contributes heavily to this factor for the Farming topic but contributed moderately to Factor 3 of the Space topic; the percentage of "to be" verbs contributes to Factor 4 for Farming and to Factor 3 for Space; fewer nominalizations contribute to Factor 4 for Space, but not for Farming (more nominalizations contributed to Factor 3 for both); and number of suggestions regarding possible diction problems contributes somewhat to Factor 4 for both topics. These data for the papers analyzed for the total sample of writers, representing the features of writing that can be used to describe performance on the two tasks, are relevant to the interpretations of the analyses for the different language groups, that are reported in the final results section.

Factor Analysis of All Variables

The data for all variables for each topic--writing and reasoning sccres assigned by human judges, WWB features extracted from the papers, and GRE part scores--were factor analyzed (Table 5 reports factor loadings of .30 and higher). The varimax factor analyses should be interpreted with caution, but can be viewed as suggestive of future possibilities. Some of the scores for the Purves/Soter and Moss scoring schemes were not highly reliable, and some of the scores do not appear to be independent of one another (holistic writing and holistic reasoning, the three Purves/Soter scores, and the Purves/Soter scores with holistic reasoning and writing), and thus could be considered to be interchangeable to some extent. However, it is interesting to note that although the holistic writing and reasoning scores had separate but high loadings on Factor 1 for the Farming topic, the loading for the holistic writing score is considerably lower than the holistic reasoning score on Factor 1 for the Space topic. Again we can observe evidence for differential performance on the two topics.

Generally, without ignoring the specific differences between topics, Factor 1 appears to reflect verbal ability, as measured both directly by the writing samples and indirectly by the GRE General Test. The Purves/Soter, Moss, and holistic writing and reasoning scores all contribute to this factor. Factor 2 for the Space topic primarily reflects GRE verbal scores, with a small contribution from the highly correlated holistic reasoning, Purves/Soter style/tone, and holistic writing scores. Factor 2 for the Farming topic, however, is dominated by WWB and Moss variables. Factors 3, 4, and 5 are formed entirely from WWB variables, which again appear to be relatively independent of the other variables and suggest that the WWB variables contribute additional information about performance on the writing tasks. In general, Factors 3, 4, and 5 reflect WWB fluency, content, and sentence variety measures, respectively.

Comparisons of Language Group Means

Regression analyses were conducted to predict the holistic writing scores and to predict part scores for the two parts of the analytical reasoning section of the GRE General Test from the full set of variables, both for the total group and for the separate language groups. Since the sample size for the Arabic language group was too small, the data for this group only contributed to the analysis of total-group data. The regression analyses should be interpreted in relation to the levels of scores obtained by the students from different native language groups (Tables 6 and 7).

The patterns of scores for the holistic writing, holistic reasoning, and Purves/Soter schemes reflect the same general trends, in which the native speakers of English received higher scores than did the Spanish speakers, who obtained higher scores than the Chinese speakers (significant at p=.001 for these F ratios for 3, 199 df). These patterns are repeated for the Moss scheme scores, except that the mean scores for the Chinese and Spanish groups are considerably lower than the qualifications score (F=16.32 for 3, 199 df at p=.001) of the English group.

The WWB scores reveal greater differentiations that do not always favor the English group. Although the English speakers committed fewer spelling errors on the Space papers (significant at the .02 level), the Chinese had fewer spelling errors than the English speakers on the Farming papers. For both the Space and Farming topics, the English papers exhibited a lower percentage of vagueness on the Farming papers (significant only at the .06 level) and the Chinese used considerably more vague words on the Space topic (p=.001). All groups, however, were relatively more similar (no significant differences) with respect to the use of abstract words and nominalizations (nominalized words) for both topics. The English papers contained considerably more (p=.001) potential problems with diction (number of suggestions) and confused homophones or word pairs (number to check) than the Chinese and Spanish papers--possibly because greater fluency is associated with greater potential for error. The English group used the passive voice most frequently in both the Space and Farming papers (p=.001 for Space, p=.04 for Farming), whereas the Chinese used it the least in the Space papers and the Spanish, in the Farming papers.

On the GRE General Test, the Chinese group scored the lowest, and the Spanish scored lower than the English speakers. The

English group received higher scores on both parts of the analytical reasoning section; but the Chinese group obtained higher scores than the Spanish group on the analytical reasoning part, and the Spanish group obtained higher scores than the Chinese group on the logical reasoning part. These scores may reflect the frequently observed higher relationship of scores on the analytical reasoning part with scores on the quantitative section and the higher relationship of scores on the logical reasoning part with scores on the verbal section. In general, Chinese nonnative speakers of English attend graduate school in the United States in mathematical or scientific fields, and the Spanish speakers tend to be more highly verbal than the Chinese, hence these patterns of relationships.

Stepwise Regression Analyses for the Total Sample and Language Groups

Stepwise regression analyses* were conducted to predict the holistic writing scores (approximately equivalent to the holistic reasoning scores) from the 21 WWB variables before further reduction to 16 variables (Tables 8 and 9). For the total group, nine variables contributed to the prediction of holistic writing scores for both Space and Farming. The two predominant variables, number of words and average word length, indicate that overall fluency contributed to the quality of the holistic evaluations. (Since the "number of words" variable was highly correlated with other WWB features, it was eliminated from the final analyses; for the Space topic, the mean number of words for the papers in the total sample was 252.27, and for the Farming topic, 212.70.)

Some relatively predictable features influenced these holistic writing scores (e.g., fewer nominalizations and spelling errors). The prediction of scores for the language groups presents a somewhat different perspective in that slightly different stylistic features contribute to the holistic scores for the different groups and for the two different topics. Since these results served to describe the performances on this particular sample of data, we cannot draw any generalizable conclusions about the features of Spanish student writing, for example, until additional studies are conducted to support the findings.

Stepwise regression analyses also were conducted to predict the part scores (analytical reasoning and logical reasoning) for the GRE General Test from the 28 score variables obtained for the Space topic and the 29 score variables obtained for the Farming

*In all regression analyses, all variables contributing to the predictions were at least moderately reliable.

topic (Table 10). For the total sample, scores on the verbal section of the GRE General Test, but not the scores on the analytical reasoning or logical reasoning parts, contributed heavily to the prediction of the logical reasoning and analytical reasoning parts, respectively. The scores from the scoring schemes contributed to a small degree (see the beta weights) to these predictions.

For the different native language groups, the analytical reasoning and logical reasoning scores were predicted predominantly by other GRE verbal scores, but scores from the scoring schemes contributed more substantially to the predictions (Tables 11-13). For the Spanish sample (Table 13), whose analytical reasoning scores were lower than those of other groups, the prediction of their analytical reasoning scores required few steps in the analyses; the prediction of their logical reasoning scores included the holistic writing scores. (Note that the Purves/Soter scores, correlated approximately in the low .70's, may be essentially interchangeable in these analyses). The stepwise regressions for the English sample (Table 12) are decidedly more prosaic--other GRE verbal scores contribute considerably to the predictions.

For the Chinese sample (Table 11), whose logical reasoning scores were lower than those of the other groups, fewer variables contributed to the prediction of their analytical reasoning scores than to that of their logical reasoning scores. Their analytical reasoning scores were not predicted by other GRE scores but instead by scores obtained from the different scoring schemes and the WWB analyses (the beta weights and correlations were very low, however). The logical reasoning scores for the Chinese sample were predicted by seven variables for each topic, dominated by GRE verbal scores, but also by scores from the scoring schemes and WWB. Again, interesting language group and across-topic differences were observed--a finding that requires replication, of course, to be generalizable.

Implications Suggested by the Moss Scheme

Because the Moss variables suggest the potential to obtain additional information about examinee reasoning skills, it is possible to speculate about the kinds of cognitive strategies and concomitant performances that are elicited by the different tasks--the types of essay topics and GRE General Test items. The two essay topics impose somewhat different demands on the writer, as follows:

- The Farming (chart/graph) topic presents data in tabular form, thus supplying a degree of structure for the relationships among the elements of the data as well as the vocabulary with which ideas can be expressed. A writer can rely extensively on these data to construct statements ("claims") regarding observed similarities and differences. Further, the writer is asked to summarize by drawing conclusions (final claims) based on the information provided. Although he or she must use cognitive strategies such as analysis and interpretation, the task is relatively concrete. In most papers, in fact, readers observed that the conclusions proposed for the Farming topic represented inferences that did not extend too far beyond the data.

- When responding to the Space (compare/contrast) topic, writers needed to draw from personal experience to develop a structure of logical relationships. The construction of advantages and disadvantages (claims) required independent analysis at a level of abstraction beyond the information given. In addition, writers were asked to render an opinion (a final claim or claims) that also required an evaluation of the similar/ differert sets of relationships they constructed. Thus, a greater range of cognitive strategies were demanded by this task. In these papers, readers observed that writers appeared to be less fluent and to exhibit more problems with vocabulary and syntax, in contrast to performance on the Farming topic. Sternglass (1986) made similar observations with respect to the different cognitive strategies called for by different types of writing tasks.

The differential task demands required by the two types of tasks, compare/contrast and chart/graph, correspond to Cummins's (1984) theoretical framework in which language tasks can be classified into four primary groups: cognitively demanding, cognitively undemanding, context-reduced, and context-embedded. The Farming topic is considerably more context-embedded than the Space topic, which is essentially more context-reduced--underscoring the different kinds or degrees of active cognitive processing required by the two tasks. Cummins's framework bears a close relationship to the results of an earlier writing assessment study (Bridgeman & Carlson, 1983, p. 52) that served as the basis for selecting the two divergent topic types. In a multidimensional scaling of the ratings made by faculty in higher education, a two-dimensional solution was obtained to reflect their perceptions of similarities and differences among topic types. The vertical axis for this solution essentially corresponded to the degree of information-processing, or complexity-of-reasoning, demands made by a topic, progressing in complexity from the top to the bottom. The horizontal axis appeared to represent the extent to which topic types demand that the writer bring personal knowledge and experience to the writing task, progressing in degree of involvement from left to right.

Thus, the types of writing tasks ranged along two dimensions, complexity and personal involvement. The chart/graph topic type fell in the upper-left quadrant of the stimulus space, whereas the compare/contrast topic type was located diagonally opposite, in the lower-right quadrant. Thus the data reported in this paper appear to confirm the judgmental ratings of the faculty respondents in the earlier survey.

Canale (1984) presents a theoretical framework to further clarify the demands of different language tasks by delineating three dimensions of language proficiency: basic, communicative, and autonomous. Autonomous language proficiency is demonstrated when focus is placed on grammatical forms, organization of ideas, and literal meaning used to convey language code and logical relationships among propositions. The previous research (Carlson & Bridgeman, 1983; Carlson et al., 1985) that served as a foundation for the present work focused primarily on communicative proficiency, particularly in the context of the writing skills of nonnative speakers of English. By taking the perspective of evaluating the reasoning processes demonstrated in essays, the current study has attempted to emphasize autonomous language proficiency to a greater degree than basic or communicative proficiency (writing skills).

Both essay topics, cf course, involve the recall and production of ideas in language, while responding to items on the GRE verbal and analytical reasoning sections requires comprehension of the task, recall of ideas, and the application of strategies for dealing with different kinds of relationships that are assessed through recognition. For the essay production tasks as well as the GRE recognition tasks, assumptions are made regarding the cognitive strategies (e.g., deduction, analysis, evaluation, critical reading) that intervene between the explicit task as it is presented and the resulting performance that is observed. Because the GRE items present different kinds of relationships, however, we can reasonably assume that all examiness have been provided with opportunities to apprehend common sets of relationships. Furthermore, because the GRE item types require an examinee to go beyond the information given as did the Space topic, a correct examinee response serves as an indicator of effective thinking skills. In the essay tasks, particularly since the instructions did not focus the writers on making their thinking explicit, however, we are not assured that their thinking is either accurately or adequately represented by the ideas produced in writing. The essay task is further confounded by the writer's engagement with the topic. As Sternglass (1986) notes, a writer's inherent cognitive potential may not be released without the personal commitment to draw upon a full range of cognitive operations leading to more complex thinking and writing.

Among the correlations (Table 3) with the part scores of the GRE verbal section, the support score for the Moss scheme yielded somewhat higher correlations (.55-.58) for the Space topic than correlations (.35-.40) with the justifications (support) score for the Farming topic. The Space topic, in fact, required the evaluation of support provided for the opinion(s) rendered, more so than was required by the Farming topic. Although the correlation was small (.18), the Moss score for graph reading skills was related to the GRE reading comprehension score and to no other GRE verbal or analytical reasoning scores, thus reflecting the demands of the Farming task.

For the Space topic, only the support score was correlated (.21) with the GRE analytical reasoning score; it was more highly correlated (.43) with the logical reasoning score. The support score represents the elaborations, justifications, and implicit and explicit opinions presented by the writer. Similarly, the logical reasoning items on the GRE General Test are intended to assess the ability to understand, analyze, and evaluate arguments, including the recognition of assumptions on which an argument is based and drawing conclusions from premises. Thus, the Moss scheme, when applied to a compare/contrast essay, represents some facets of reasoning that the logical reasoning test is intended to measure--predominantly support for an argument but also the claims and qualifications making up the argument.

For the Farming topic, the Moss scores were reflected equally in correlations among GRE analytical reasoning and logical reasoning scores with deductions, claims, and justifications. The correlations of the Moss qualifications score with the logical reasoning score, however, were moderately high (.42) for the Farming topic and low (.22) for the Space topic, but they were not significant with the analytical reasoning scores. The Farming topic apparently elicited more statements to qualify the claims that could be made about the data, and justifiably so. Since writers were required to create their own structures of relationships for the Space topic, though, their claims were less likely to be stated if they required qualification. Here again, the higher degree of evaluation required by the Space topic is suggested. These correlations of Moss qualifications scores with GRE logical reasoning scores possibly indicate the degree to which the logical reasoning items focus on evaluation through critical analysis or critical reading.

In the varimax factor analysis of all variables (Table 5), the Moss support (justifications), claims, and qualifications scores loaded on one factor with the GRE verbal and both the logical reasoning and analytical reasoning scores for the Farming topic. For the Space topic, the GRE analytical reasoning score contributed to Factor 2, in which the logical reasoning and GRE verbal scores loaded most highly. Again, performance on the compare/contrast Space topic suggests the separate dimensions of reasoning tapped by logical reasoning and analytical reasoning.

Connor (1987) and Connor and Lauer (1987) have used a different adaptation of the Toulmin approach for evaluating the effectiveness of persuasion in compare/contrast essays. Their scheme, in contrast with the more analytical, "frequency count" approach of the Moss scheme, involves ratings of three elements of a paper--claims, data, and warrants. In one analysis, using a holistic (writing) rating as the dependent score, 48 percent of the variance was contributed by the Informal Reasoning score obtained by using their version of the Toulmin model. In another study, the three Informal Reasoning scores (claims, data, warrants) were differentially correlated with the holistic scores for compare/contrast essays; the claim score yielded the highest correlation (.72) with the holistic scores. Since these scores were derived differently than scores for the Moss scheme, direct comparisons cannot be made. However, their research suggests that some form of a Toulmin scheme may provide meaningful information about the reasoning strategies used in written persuasion.

The analyses in the present study have demonstrated that the scores derived for the Moss scheme appear to reflect separable components of an argument, particularly for essay topics (compare/contrast) that require the construction of structure and evaluation of support. Because these skills need to be demonstrated in an essay rather than applied when responding to a recognition item, they also may be influenced by the writer's training (e.g., in formal logic), stylistic preferences in responding to the task, and/or degree of engagement with the topic. Such variables that influence performance on a production task would need to be controlled (e.g., with explicit instructions indicating how papers would be evaluated) in order to provide examinees with equivalent opportunities to demonstrate their abilities optimally.

IV. SUMMARY OF RESULTS

This investigation demonstrates the value of comparing observations yielded by several different instruments, and using a variety of statistical approaches (descriptive statistics, correlations, factor analyses, regression analyses) in order to better understand the interrelationships of the features of written discourse to thinking and writing skills with subscores derived from the verbal and analytical sections of the GRE General Test. The limitations of the study do not permit us to generalize these observations too broadly, but the findings can serve as a basis for further work. Overall, we observed the following:

- The reliability of scoring schemes applied by human judges can vary considerably, depending on the features of writing the readers are able to identify in samples in which actual performance varies with the task and native language group. Some of the more important features of writing and thinking skills are the most difficult to identify with reliability, perhaps because they are less frequently demonstrated in papers and are more difficult to define for reader consensus when samples of writing are being evaluated. In addition, these skills appeared to be more difficult to recognize and identify because writers used a variety of techniques to express their ideas in language. In some papers, for example, ideas could be readily apprehended by the surface structure of the organization, syntax, and vocabulary in which they were stated. In other papers, though, ideas were communicated at a level in which the deep structure predominated in the making of meaning. Readers who applied the Moss scheme commented on this phenomenon, and possibly were more alert to such nuances of language because of the kinds of variables on which this scheme was focused. Subsequently, they found it more difficult to accurately identify certain reasoning processes when they were not immediately obvious. The differences in performance on the two tasks also might have contributed to reduced reliability, since variables identified in one kind of task would differ in terms of the ways ideas were expressed in order to meet the demands of the task.

- Scores based on the Purves/Soter scheme, despite the strenuous attempts of the developers and readers to separate scores from writing skills, were highly related to, therefore confounded by, verbal ability.

- The Moss scheme, in which verbal fluency was extracted from the scores, attempted to identify some higher-order reasoning skills that appeared to occur infrequently, thus leading to low estimates of reader reliability. Some of the more reliable scores, however, appeared to be independent of verbal ability, yet were not highly related to the part scores for the analytical reasoning section of the GRE General Test. It is possible that these scores, particularly those for the Space topic, tapped reasoning skills that are not independently assessed by the GRE. Further research is needed to determine what these scores mean and whether they represent important developed abilities.

o The Writer's Workbench provides information about the characteristics of text as well as measures that identify features of written discourse that are relatively independent. The WWB has the potential to supply diagnostic feedback in writing instruction, as is being pursued at Colorado State University and elsewhere. It might also serve, with more research, as a tool for equating prompts for open-ended assessment by generating evidence that the prompts elicit the same features and ranges of performance.

o The data support the contention that different task demands elicit different kinds of writing performance, and that the writing performance of students representing different first-language groups varies in complex ways in response to these tasks as well. Park (1986), analyzing the data from this study, also confirmed these differential topic effects. He used scores from the Purves/Soter scheme and other measures of textual features, and concluded that significant topic effects were found in elaboration-length measures, the Purves/Soter content/thinking scores, and the holistic scores. Topic effects were not significant, however, for syntactic complexity measures and the other Purves/Soter scores. Thus, three major results pertinent to the effects of topic were obtained: (1) writing on the Space writing task evidenced more elaboration (longer essay) and higher quality on the content/thinking score than was evidenced in the Farming task; (2) writing on both tasks exhibited the same degree of syntactic complexity and quality in organization and style/tone (Purves/Soter) scores; and (3) writing on the Farming task yielded higher holistic scores.

Relationships of Reasoning Scores to GRE General Item Types

Descriptive statistics

Native speakers of English obtained higher mean scores on the item types in the verbal and analytical reasoning sections of the GRE General Test than did students in the Chinese and Spanish samples (Table 7). Mean scores on the GRE item types were higher for the Spanish sample, except for the analytical reasoning item type, for which the Chinese mean score was higher. Since Chinese students tend to pursue graduate studies in scientific and quantitative fields, and since the analytical reasoning scores tend to be more highly correlated with the quantitative section of the GRE, this trend may be repeated in these data.

Correlational analyses

For this relatively small sample of students, predictable intercorrelations among scores for the different item types on the GRE General Test were obtained (Table 3). Scores on the logical reasoning item types were more highly correlated with scores on the verbal item types than were scores on the analytical reasoning item types. The correlation between analytical reasoning and logical reasoning scores was only moderate (.48).

For the scores obtained from the Purves/Soter scheme, the same pattern and levels of relationships were observed. The Purves/Soter scores were all highly correlated with one another, particularly within topics (.88-.93), and also across topics (.72-.77). The correlations of these scores with GRE verbal scores were relatively high as well (.59-.72), and were more highly correlated with GRE logical reasoning scores (.59-.65) than with GRE analytical reasoning scores. Thus, the scores yielded by the Purves/Soter scheme, which strongly reflect verbal ability, do not appear to contribute information beyond that which is obtained from scores on the GRE General Test. The high correlations (.75-.82) of Purves/Soter scores with holistic writing and reasoning scores further support this conclusion.

The holistic reasoning and writing schemes for both topics reflect the same pattern and levels of relationships to scores on the GRE General Test as were obtained for the Purves/Soter scores. Again, the holistic scores were more highly correlated with GRE logical reasoning scores than with GRE analytical reasoning scores, and had relatively high correlations (.69-.76) with GRE scores on the three verbal item types.

These patterns of relationships among the GRE verbal, analytical reasoning, logical reasoning, Purves/Soter, and holistic writing and reasoning scores provide some evidence for convergent validity, particularly supporting the contributions of "general" verbal skills to the GRE scores in the analyses. The data suggest, in fact, that the kinds of criterion measures selected for validity analyses, as well as the ways in which they are scored, influence the kinds of information that can be obtained regarding the validity of tasks presented on the GRE General Test. The Purves/Soter and holistic writing and reasoning scores were not differentially correlated with the separate GRE verbal scores, or with the GRE verbal scores in relation to the GRE logical reasoning scores. The GRE analytical reasoning score, in which performance appears to be less influenced by general verbal skills, retains its relative independence from other GRE scores when contrasted to certain criterion measures.

Scores on the Moss scheme, however, present a different picture of relationships, although refinement of the Moss scheme, new data, and subsequent replication would be required for assurance that these differences exist. The correlations of scores on the Moss scheme with one another and across topics generally are low or insignificant. Within the Space topic, the claims and support scores were moderately correlated (.55); within the Farming topic, the justifications and claims scores were more highly correlated (.64). Thus, in general, the Moss scheme scores tended to be more independent.

Correlations of Moss scores with holistic reasoning scores were somewhat higher than correlations with holistic writing scores, although the support scores on the Space topic yielded correlations of approximately the same magnitude (.59-.77) for holistic writing and reasoning scores on both topics. Correlations of Moss scores with verbal and analytical reasoning item type scores on the GRE General Test ranged from low to moderate. Fewer Moss scores were significantly correlated with GRE analytical reasoning scores, but the usually clear differences between analytical reasoning and logical reasoning item types were not obtained. The Moss scores on the Space topic were somewhat more highly correlated with scores on the three GRE verbal item types than were scores on the Farming topic. The latter finding may be attributed to differences in task demands of the two topics that were reflected by the design of the Moss scoring schemes for Space and Farming and that were made more apparent through the Writer's Workbench analyses (independent of the scoring schemes). Until further investigation, we can conclude that, for these data, the Moss scheme scores are related to, but relatively independent of, scores on item types of the GRE General Test. Additional data would need to be collected and analyzed to determine whether this independence can be observed in different samples, and whether the scores provide meaningful information about reasoning skills.

Factor analyses

When all variables were factor analyzed for each topic, scores on the item types of the GRE General Test contributed differentially to the factor loadings (Table 5). For the Farming topic, all GRE scores on item types in the verbal and analytical reasoning sections loaded on Factor 1, although the loading for analytical reasoning (.44) was considerably lower than the loadings for logical reasoning and the verbal item types. This result reflects the observed correlations and the pattern typically observed for analytical reasoning scores in relation to scores on verbal measures. The holistic writing scores and several of the reasoning scores yielded by the scoring schemes contributed approximately equivalent loadings to Factor 1, as did scores on the GRE verbal item types. (Although the holistic writing and reasoning scores were highly correlated, they appear to provide somewhat different information in these analyses). The other factors obtained for the Farming topic were not influenced by GRE scores, but primarily by Writer's Workbench features. Factor 2, however, included moderately high loadings for scores derived from the Moss scheme.

In the factor analysis of all variables for the Space topic, GRE scores for the logical reasoning item type and the three verbal item types contributed substantially to Factor 2, with the analytical reasoning score loading moderately (.44). The GRE scores, excepting the analytical reasoning score, also loaded, though moderately, on Factor 1. This factor is primarily composed of scores yielded by the scoring schemes. It contains a high loading or holistic reasoning, which had the highest loading on Factor 1 for the Farming topic. Just as for the Farming topic, the three remaining factors are constituted by approximately equivalent Writer's Workbench features.

The factor analyses provide information regarding the potential contributions of different verbal measures in a measurement domain that represents general verbal ability and evaluations of verbal writing and reasoning performance. Clearly, a general verbal ability component predominates, yet the analytical reasoning score on the GRE General Test retains some independence from the logical reasoning score, whereas the logical reasoning score is more highly related to other general verbal ability measures. The scores obtained from the various judgmental schemes appear to provide additional information, although limited. The contributions of the Moss scheme scores to Factor 2 for the Farming topic suggest that the task demands of this chart/graph topic type may provide relatively independent information, particularly associated with concreteness (versus the negative loading for vagueness). (The loading for prepositions serves as an indicator of a kind of performance required by the task of summarizing results displayed on a graph.) The other factors (3-5) constitute features that would be reflected in open-ended writing tasks, quite separate from performance on recognition measures.

Regression analyses

All variables were analyzed to predict scores on the two analytical reasoning item types, analytical reasoning and logical reasoning, for the total sample (Table 10). For both topics, the GRE reading comprehension and sentence completion scores dominate in the prediction of logical reasoning scores. The holistic scores and a few Writer's Workbench and Moss scores contribute relatively small amounts of information to the prediction of these scores. The crosscultural differences observed in regression analyses (Tables 11-13) are described in the following section.

Within the perspectives of these data, the GRE General Test scores appear to provide meaningful information about verbal ability that is consistent with relationships found in other GRE validity studies. The factor analyses for the total sample and analyses for the language groups suggest that additional information about verbal ability might be obtained, but these data do not indicate whether that information would be meaningful or important. As observed in the previous GRE/TOEFL writing assessment study, performance on open-ended production tasks differs somewhat from performance on recognition tasks. Evaluations of these tasks are further confounded by the observed differences in performance in response to different tasks such as the Space and Farming topics. These differences were not apparent in the correlations among variables or in differential correlations with the GRE analytical and logical reasoning scores, but were revealed particularly by Writer's Workbench features in the language group regression analyses. Given a carefully designed chart/graph task, or one that requires specific yet significant reasoning skills, it may be possible to obtain information about these skills that is not yielded by the analytical reasoning section of the GRE General Test.

These speculations revolve full circle to the Formulating Hypotheses (FH) item type, referred to in the introduction, which has more definitively provided independent, potentially significant information regarding hypothesis-generation skills. The tasks presented in the FH problems closely resemble the chart/graph writing task, but constrain the responses to discrete, short answers. A complementary study (Carlson & Ward, in preparation) proposes the design of a computer system to deliver the FH item types that, as a preliminary prototype, would provide data regarding the design of assessment tasks that indeed might yield significant information about examinee reasoning skills that cannot be obtained through recognition measures.

CROSSCULTURAL DIFFERENCES

Overall, the mean scores for the four scoring schemes and GRE General Test part scores were higher for the English group than for the Spanish group, whose scores were higher than for those of the Chinese group. The WWB programs provided greater differentiation among the language groups, since papers written by each of the groups could be described by quite different combinations of Workbench features.

For the Space topic, the Chinese papers contained more vague words but less passive voice and fewer potential usage problems, short sentences, "to be" verbs, content words, and prepositions. The Spanish papers for the Space topic presented more spelling errors, somewhat more sentence variety, and fewer adjectives. The English papers contained more potential usage problems, "to be" verbs, passive voice, content words, and adjectives.

For the Farming topic, the Chinese papers presented fewer potential usage problems, spelling errors (than the Spanish), and prepositions, and less sentence variety. The Spanish papers exhibited more spelling errors, more vague words (than the English), greater sentence variety, and less passive voice. The English papers for the Farming topic contained considerably more diction problems, "to be" verbs, passive voice, content words, and adjectives.

For both topics, all three language groups appeared to be relatively similar with respect to number of nominalizations, word length, and abstractness. The native speakers of English used the passive voice more frequently in their Space and Farming papers, and also exhibited more instances in which diction and confused homophones or word pairs may have been problematic. Perhaps the greater fluency of the English students resulted in more variety and risk-taking, whereas nonnative speakers of English tended to constrain their writing by using "safe" syntax and vocabulary. The observed mean differences were reinforced by the stepwise regression analyses, which indicated that different variables contributed to the prediction of holistic writing (and reasoning) scores and scores on the two parts of the GRE analytical reasoning sections for the three language groups. Somewhat different variables contributed to the predictions for the two different writing tasks as well.

Again it should be noted that, with the exception of reliability estimates for the holistic reasoning scores, estimates of reader reliability for the three scoring schemes that required human judgments differed, unsystematically and inexplicably, for the separate language groups. These differential reliability estimates, however, should not be viewed as generalized trends without further investigation. These data suggest that, when working with judgmental data involving students from different cultures, the reliability of judgments within each language group sample should be determined, since they might be indicative of some form of bias.

Based primarily on computerized analyses of textual features, we have observed that speakers whose primary languages are other than English use different language devices to express their ideas when writing in English. Stepwise regression analyses predicting their holistic writing scores suggest that different features of writing combine to achieve "successful" papers. For the total sample, for example, the number of words and average word length contributed substantially to the holistic ratings for both topics (Tables 8 and 9). For the Chinese sample, however, the use of concrete and explicit (vs. abstract and vague) words and shorter (vs. longer) sentences on the Space topic influenced their holistic scores, whereas the use of prepositions was dominant for the Farming topic. In contrast, the holistic scores for the Spanish sample were predicted by the use of passive voice and shorter sentences on the Space topic, and by the number of sentences and avoidance of potential errors (fewer suggestions) on the Farming topic.

Analytical reasoning and logical reasoning scores on the GRE General Test were predicted by different combinations of variables for these different language groups (Tables 11-13). For the English sample, the pattern of predictions resembled the predictions of these scores for the total sample (Table 10), for which scores on the reading comprehension item types substantially contributed to the prediction of analytical reasoning scores. Scores on the reading comprehension item types and, to a lesser degree, on the sentence completion item types predominantly predicted logical reasoning scores for the total sample. In contrast, scores on GRE item types did not contribute to the prediction of analytical reasoning scores for the Chinese sample, although sentence completion and reading comprehension scores contributed to the prediction of logical reasoning scores (Table 11). For the Space topic, the holistic reasoning score, though with a low beta weight (.20), predicted the analytical reasoning score; for the Farming topic, the Moss claims and graph reading scores predicted the analytical reasoning score. Thus, analytical reasoning scores were predicted by scores from the reasoning schemes for the Chinese sample, but not for the English and Spanish samples.

These findings of crosscultural differences are not surprising, placed within the perspectives of the knowledge gained in the fields of writing assessment and crosscultural rhetoric. It is important to recognize that competent writing is a construct that requires careful definition in order to be measured. The definition of this construct may vary from occasion to occasion. Since competent writing is situationally dependent, it is defined by the specific task demands within the particular situation in which, and for which, writing ability is being assessed. The evaluation of writing samples produced by international students writing in English is made more challenging by the recognition that, as task demands vary, so do the forms of responding to the task.

The writing skills demonstrated by a writer are influenced by the complex interactions of the writer and demands of a task within a sociocultural context. Any writing assessment, however formal or informal, is conducted for a social purpose, such as to provide instructional feedback, determine course placement, or enforce standards for admission to or exit from educational training or occupations. Because effective writing can take a variety of forms to communicate ideas, the features of a piece, given the same purpose and audience, can vary considerably from writer to writer. The assessment of skills embodied in writing involves a social situation in which judgments about performance and subsequent interpretations extend beyond the writer and reader. In responding to this social situation, writers must attempt to meet the expectations of readers as they simultaneously bring their own expectations to the task. Making a match between reader and writer expectations involves yet another complex interaction, since readers and writers have acquired their idiosyncratic approaches to defining competent performance largely through educational experiences within their particular cultures.

This research has not attempted to explain the reasons for the expected cultural differences that were obtained, but instead has documented further the cultural differences that can be elicited in response to different assessment tasks by applying different approaches to perceiving and evaluating performance. Extensive research in contrastive rhetoric involves differences of opinion regarding the ways in which language and though are culturally shaped (Kaplan, 1982, 1987). Experts in contrastive rhetoric generally agree, however, that the acculturation process of learning to write takes place in schools as students learn to write according to certain conventions. These cultural conventions may indeed have contributed to the shaping of thought in a particular culture (Indrasuta, 1987; Kadar-Fulop, 1987). Most important, though, is the observation that students have learned to write in ways that are bound by conventions that affect organization, syntax, and vocabulary choice (Purves, 1987). Since these writing data were collected at the point of entry to graduate institutions, it was necessary to recognize that writers coming from different cultures have learned rhetorical patterns that may differ from patterns used in the United States, rhetorical patterns that were reinforced by their educational experiences in their specific cultures. Thus, the writing and reasoning performance of nonnative speakers as well as native speakers of English could not be evaluated in terms of traditional, possibly narrow, views of appropriate forms or features of written products. This research suggests approaches to our understanding of the intricacies that confound thought and written language.

With refined instruments, such as an adaptation of the Moss scheme, additional writing samples, and a larger sample size, it would be interesting to investigate further the potential crosscultural differences in responding to writing and reasoning tasks in both recognition and open-ended formats. Furthermore, if systematic differences are observed, it would be important to investigate the extent to which potential cultural differences in responding might influence performance on the GRE General Test and in graduate school.

Conclusions

Several experts in the field contributed considerable effort in an attempt to identify significant features that might serve as indicators of effective reasoning skills in written discourse. This kind of exploration, despite its uncontrollable limitations, is essential to this line of inquiry. What we have learned, it is hoped, will contribute to further extensions of some of the more promising directions that were pursued:

o Different communities of readers who can generally agree on the features of writing that evidence effective reasoning skills bring different approaches, labels, and definitions to the observation of these skills. The interrelationships within and among the different scoring schemes provide information about what we might be evaluating as well as the degree to which these skills are evaluated independently of other skills with which they are, perhaps, inextricably confounded. We need to collect more data to determine to what extent performance differences in text are a function of tasks and scoring methods, or of actual differences in developed abilities.

o The reasoning skills that are deemed to represent important, high-level abilities are difficult to identify with reliability and validity in measures that require open-ended production. One of the reasons for this difficulty is that such higher-order skills did not appear as frequently as would be expected in the papers written by graduate-level students. This may have been a function of the writing task demands and/or might be observed in student writing in general.

- In addition to the differential performance that is elicited in response to different task demands, students who have been trained in academic writing in different cultures may have different perceptions of performance expectations. Within the context of academic writing in the United States, it is possible that nonnative speakers of English, in the early stages of adapting to our expectations, still reflect strong dispositions to write as they were trained to write (and think) in their cultures. At some intermediate point in their educational experiences in the United States, their writing may exhibit a combination of cultural influences that vary in effectiveness. Finally, we might expect that academic training in the United States would lead toward more Westernized approaches, with a caveat--against what standards is it appropriate to evaluate effective academic writing? Furthermore, against what standards is it appropriate to evaluate the quality of a student's reasoning skills when expressed through the vehicle of writing? These standards are not sufficiently explicit and vary, depending on the contexts--major fields, types of writing tasks, and whether or not the students who return to their native countries will need to make accommodations when writing in English within their native cultures.

These observations lead to an even more basic question: do the features observed in written discourse contribute differentially to the quality of writing, depending on the different types and contexts of writing, and are any of these features generalizable across types and contexts? Thus, we continue to face and investigate a challenging dilemma, to clarify and communicate consistently the expectations of readers, writers, and interpreters of evaluations of written discourse.

V. REFERENCES

American Psychological Association. (1985, February). Joint technical standards for educational and psychological testing. Washington, DC: American Psychological Association.

Breland, H. M., & Jones, R. J. (1982). Perceptions of writing skill. (College Board Report No. ˜82-4, and ETS RR No. 82-47). New York: College Entrance Examination Board.

Bridgeman, B., & Carlson, S. (1983). Survey of academic writing tasks required of graduate and undergraduate foreign students. (ETS Research Rep. No. 1983-18). Princeton, NJ: Educational Testing Service.

Canale, M. (1984). On some theoretical frameworks for language proficiency. In C. Rivera (Ed.), Language proficiency and academic achievement (pp. 28-40). Clavedon, England: Multilingual Matters, Ltd.

Carlson, S. B., Bridgeman, B., Camp, R., & Waanders, J. (1985). Relationship of admission test scores to writing performance of native and nonnative speakers of English. (TOEFL Research Rep. No. 19). Princeton, NJ: Educational Testing Service.

Carlson, S. B., & Bridgeman, B. (1986). Testing ESL student writers. In K. L. Greenberg, H. Wiener, & R. A. Donovan (Eds.), Writing assessment: Issues and strategies. New York: Longman.

Carroll, J. (1960). Vectors of prose style. In T. A. Sebeok (Ed.), Style in language (pp. 283-292). Cambridge, MA and NY: Technology Press and John Wiley.

Cherry, L. L., Fox, M. L., Frase, L. T., Gingrich, P. S., Keenan, S. A., & Macdonald, N. H. (1983, May/June). Computer aids for test analysis. Bell Laboratories Record.

Connor, U. (1987). Personal communication.

Connor, U., & Lauer, J. (1987, in press). Crosscultural variation in persuasive student writing. In A. C. Purves (Ed.), Written communications annual, Vol. 2. Newbury Park, CA: Sage.

Cummins, J. (1984). Wanted: A theoretical framework for relating language proficiency to academic achievement among bilingual students. In C. Rivera (Ed.), Language proficiency and academic achievement (pp. 2-27). Clavedon, England: Multilingual Matters, Ltd.

Educational Testing Service. (1986). GRE guide to the use of the Graduate Record Examinations program, 1986-87. Princeton, NJ: Author.

Frase, L. T. (in press). Computer analysis of written materials. In D. Reinking (Ed.), Computers and reading: Issues for theory and practice. New York: Teachers College Press.

Frase, L. T., Kiefer, K. E., Smith, C. R., & Fox, Mary L. (1985). Theory and practice in computer-aided composition. In S. W. Freedman (Ed.), The acquisition of written language. Norwood, NJ: Ablex.

Indrasuta, C. (in press). Contrastive rhetoric and the relationship among culture, language and thought. In A. C. Purves (Ed.), Written communications annual, Vol. 2. Newbury Park, CA: Sage.

Kadar-Fulop, J. (in press). Culture, writing, and the curriculum. In A. C. Purves (Ed.), Written communications annual, Vol. 2. Newbury Park, CA: Sage.

Kaplan, R. B. (in press). Contrastive rhetoric and second language learning. In A. C. Purves (Ed.), Written communications annual, Vol. 2. Newbury Park, CA: Sage.

Kaplan, R. B. (1982). Contrastive rhetoric: Some implications for the writing process. In I. Pringle, A. Freedman, & J. Yalden (Eds.), Learning to write: First language, second language. London: Longman.

Khoury, B. (1984, March). Future of the GRE analytical measure. Report presented at the meeting of the Graduate Record Examinations Board, Savannah, GA.

Kiefer, K. E., & Smith, C. R. (1984). Textual analysis with computers: Tests of Bell Laboratories' computer software. Research in the Teaching of English, 17(3), 201-214.

Moss, P. (1986). Personal communication.

Park, Y. M. (1986). The influence of task upon writing performance. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Purves, A. C. (Ed.). (in press). Written communications annual, Vol. 2 (preface). Newbury Park, CA: Sage.

Purves, A. C. (1986). Rhetorical communities, the international student, and basic writing. Journal of Basic Writing, 5(1), 38-51.

Purves, A. C. (1985). Framework for scoring: GRE/TOEFL. Unpublished manuscript, University of Illinois at Urbana-Champaign, Curriculum Laboratory.

Purves, A. C., Soter, A., Takala, S., & Vahapassi, A. (1984). Towards a domain-referenced system for classifying composition assignments. Research in the Teaching of English, 18, 385-409.

Reid, S., & Findlay, G. (1986). Writer's Workbench analysis of holistically scored essays. Computers and Composition, 3(2), 6-32.

Ruth, L., & Murphy, S. (1987). Designing writing tasks for the assessment of writing. New York: Ablex.

Smith, C. R., & Kiefer, K. (1982, April). Writer's Workbench: Computers and writing instruction. Paper presented at the Proceedings of the Future of Literacy Conference, University of Maryland, Baltimore.

Soter, A. (in press). The second language learner and cultural transfer in narration. In A. C. Purves (Ed.), Written communications annual, Vol. 2. Newbury Park, CA: Sage.

Soter, A. (1985). GRE/TOEFL scoring criteria. Unpublished manuscript, University of Illinois at Urbana-Champaign, Curriculum Laboratory.

Sternglass, M. (1986). Commitment to writing and complexity of thinking. Journal of Basic Writing, 5(1), 77-88.

Toulmin, S., Rieke, R., & Janik, A. (1984). An introduction to reasoning, (2nd ed.). New York: Macmillan.

Ward, W. C., Frederiksen, N., & Carlson, S. (1980). Construct validity of free-response and machine-scorable versions of a test. Journal of Educational Measurement, 17(1), 11-29.

Wilson, K. M. (1985). The relationship of GRE General Test item-type part scores to undergraduate grades. (GRE Professional Report, GRE No. 81-22P). Princeton, NJ: Educational Testing Service.

Table 1

Correlations of Holistic Scores Total* and GRE Item Type Scores

(sample of 132 cases)

				GRE Scores					
			Verbal			Quantitative		Analytical	
Scores	Hol.	SC	DV	RC	QC	M	DI	AR	LR
GRE Verbal									
Sentence Completion (SC)	.68								
Discrete Verbal (DV)	.67	.64							
Reading Comprehension (RC)	.70	.70	.64						
GRE Quantitative									
Quantitative Comparisons (QC)	-.22	-.26	-.30	-.12					
Discrete Math (M)	-.31	-.28	-.36	-.26	.76				
Data Interpretation (DI)	-.09	-.03	-.08	.00	.64	.59			
GRE Analytical									
Analytical Reasoning (AR)	.23	.15	.17	.24	.46	.35	.50		
Logical Reasoning (LR)	.64	.65	.50	.67	-.09	-.18	.02	.24	

*Holistic scores averaged over four writing samples

Table 2

Means and Standard Deviations of Variables in Analysis
(N=203)

	Space Topic		Farming Topic	
	(29 variables)		(30 variables)	
Variables	Mean	S.D.	Mean	S.D.
Holistic Writing	4.23	1.57	4.34	1.58
Holistic Reasoning	3.49	1.21	3.53	1.33
Purves/Soter Scheme				
Content	18.44	4.80	17.72	4.94
Organization	8.23	2.14	8.01	2.50
Style and Tone	7.56	2.12	7.70	1.99
Moss Scheme-Space				
Claims	5.88	2.20	--	--
Support	2.67	1.53	--	--
Qualifications	.45	.54	--	--
Moss Scheme--Farming				
Graph Reading	--	--	3.70	1.36
Deductions	--	--	.89	1.00
Claims	--	--	2.70	1.21
Justifications	--	--	1.63	1.13
Qualifications	--	--	.15	.25
Writer's Workbench Variables				
Number of suggestions	6.14	4.58	6.35	5.06
No. spelling errors	2.81	2.55	2.09	2.47
% vague words	6.71	2.54	5.22	2.67
Number to check	2.50	1.77	2.07	1.66
Avg. sentence length	21.10	6.69	22.11	7.50
% shorter sentences	27.40	14.49	28.12	13.34
% longer sentences	11.58	8.40	11.42	9.25
% "to be" verbs	81.14	11.86	75.21	18.74
% passive voice	11.32	9.09	11.56	10.46
% nominalizations	3.56	1.42	2.80	1.32
Avg. word length	4.70	.31	4.65	.27
% content words	54.88	3.80	57.48	4.00
Avg. length content words	6.20	.46	5.86	.67
% prepositions	11.00	2.42	13.33	3.35
% adjectives	15.07	3.14	17.60	3.60
% abstract	3.55	1.44	2.90	1.32

GRE General Test Scores	Mean	S.D.
Verbal		
Sentence completion	6.99	3.85
Discrete verbal	19.02	8.89
Reading comprehension	11.95	5.60
Analytical Reasoning		
Analytical reasoning	21.25	6.61
Logical reasoning	5.85	2.77

Table 3

Correlations of Scoring Schemes and GRE Part Scores (N=203)
(.001 level of significance)

			GRE General Test					Purves/Soter Scheme						Moss Scheme								Holistic Schemes			
								Space			Farming			Space			Farming					HR		HW	
			SC	DV	RC	AR	LR	CT	O	ST	CT	O	ST	C	S	Q	G	D	C	J	Q	S	F	S	F
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
GRE V	1	Sent Comp																							
	2	Disc V	.81																						
	3	Read Comp	.80	.77																					
GRE AR	4	Analyt R	.39	.39	.43																				
	5	Logical R	.76	.68	.77	.48																			
Purves/	6	Cont/Think	.67	.66	.72	.33	.62																		
Soter	7	Organization	.69	.66	.71	.29	.62	.93																	
Space	8	Style & Tone	.70	.69	.72	.33	.65	.91	.90																
Purves/	9	Cont/Think	.64	.61	.68	.33	.59	.73	.74	.73															
Soter	10	Organization	.62	.59	.66	.34	.61	.72	.73	.72	.91														
Farming	11	Style & Tone	.65	.64	.69	.34	.64	.76	.76	.77	.88	.88													
Moss	12	Claims	.31	.35	.35		.28	.52	.52	.52	.37	.33	.36												
Space	13	Support	.55	.57	.58	.21	.43	.72	.71	.71	.58	.56	.57	.55											
	14	Qualification	.32	.38	.35		.22	.32	.33	.35	.31	.30	.30	.17*	.32										
Moss	15	Graph			.18*			.26	.25	.24	.22	.31	.28		.26										
Farming	16	Deduction	.25	.18*	.28	.22	.26	.24	.24	.21	.17*	.24	.26			.23	.19								
	17	Claims	.40	.46	.42	.35	.37	.41	.40	.41	.51	.45	.39	.22	.30										
	18	Justification	.35	.36	.40	.20	.30	.37	.38	.39	.57	.51	.41	.22	.31				.64						
	19	Qualification	.44	.41	.38		.42	.42	.42	.39	.38	.41	.48	.18*	.31	.28	.17*		.19*						
Holistic	20	Space	.68	.72	.70	.37	.60	.81	.81	.79	.69	.70	.70	.63	.77	.42	.24	.24	.43	.39	.41				
Reasoning	21	Farming	.72	.69	.74	.41	.64	.74	.75	.75	.81	.80	.78	.35	.59	.34	.33	.25	.64	.61	.47	.74			
Holistic	22	Space	.76	.74	.76	.35	.70	.82	.81	.80	.70	.71	.74	.44	.65	.34	.23	.27	.37	.34	.43	.80	.74		
Writing	23	Farming	.72	.71	.76	.37	.69	.78	.76	.78	.78	.77	.79	.45	.60	.34	.30	.26	.48	.46	.39	.76	.82	.85	

*at .01 level.

Table 4

Varimax Factor Analysis of Writer's Workbench Variables
(N = 203)

Farming Topic		Space Topic	
Factor 1	Loading	Factor 1	Loading
% content words	.99	% content words	.99
% number of adjectives	.65	% adjectives	.69
Number of spelling errors	-.22		
Avg. word length	.40		
Factor 2		Factor 2	
% shorter sentences	.79	% shorter sentences	.88
Avg. sentence length	.60	Avg. sentence length	.61
% longer sentences	.52	% longer sentences	.53
Number to check	.29	Number of spelling errors	.32
% prepositions	.25	Avg. word length	.55
Factor 3		Factor 3	
Avg. word length	.76	Avg. length content words	.89
Avg. length content words	.75	Avg. word length	.79
% nominalizations	.37	% prepositions	.62
% vague words	-.35	% vague words	-.59
% prepositions	.26	% passive voice	.56
		% "to be" verbs	.47
		Number to check	.35
		Avg. sentence length	.34
		% nominalizations	.31
Factor 4		Factor 4	
% passive voice	.71	% nominalizations	-.62
% "to be" verbs	.37	Number of suggestions	.44
Number of suggestions	.30	% abstract	-.15
% abstract	-.14	Number to check	.35
(accounting for 35.5% variance)		(accounting for 47.9% variance)	

Table 5

Varimax Factor Analysis of All Variables
(N = 203)

Farming Topic		Space Topic	
Factor 1	Loading	Factor 1	Loading
Holistic reasoning	.92	Purves/Soter content	.86
Purves/Soter content	.89	Purves/Soter organization	.84
Holistic writing	.87	Holistic reasoning	.82
Purves/Soter organization	.86	Purves/Soter style & tone	.82
Purves/Soter style & tone	.85	Moss support	.78
GRE reading comprehension	.82	Holistic writing	.68
GRE sentence completion	.79	Moss claims	.63
GRE discrete verbal	.77	WWB number to check	.48
GRE logical reasoning	.74	WWB number of suggestions	.41
Moss justification	.63	Moss qualifications	.31
Moss claims	.62	GRE logical reasoning	.42
WWB number of suggestions	.54	GRE sentence completion	.50
GRE analytical reasoning	.44	GRE reading comprehension	.54
WWB number to check	.44	GRE discrete verbal	.51
Moss qualifications	.43		
WWB % "to be" verbs	.22		
WWB % abstract	-.15		
WWB avg. word length	.36		
Factor 2		Factor 2	
WWB % prepositions	.58	GRE logical reasoning	.71
WWB % vague words	-.56	GRE sentence completion	.70
Moss deductions	.40	GRE reading comprehension	.63
Moss graph reading	.39	GRE discrete verbal	.62
Moss claims	-.45	GRE analytical reasoning	.44
Moss justifications	-.33	Holistic reasoning	.30
		Purves/Soter style & tone	.34
		Holistic writing	.48
Factor 3		Factor 3	
WWB avg. word length	.84	WWB avg. length content words	.83
WWB avg. length content words	.71	WWB avg. word length	.72
WWB % nominalizations	.34	WWB % prepositions	.50
		WWB % vague words	-.49
		WWB % nominalizations	.47
		WWB % passives	.46
		WWB % "to be" verbs	.42
Factor 4		Factor 4	
WWB % content words	.95	WWB % content words	.88
WWB % adjectives	.64	WWB % adjectives	.71
WWB % passive voice	-.18	WWB avg. word length	.52
		WWB % nominalizations	.34
Factor 5		Factor 5	
WWB % shorter sentences	.78	WWB % shorter sentences	.88
WWB avg. sentence length	.56	WWB avg. sentence length	.61
WWB % of longer sentences	.54	WWB % longer sentences	.59
WWB number of spelling errors	.26	WWB number of spelling errors	.28
		WWB % abstract	-.08
(accounting for 48.4% variance)		(accounting for 56.0% variance)	

Table 6

Means and Standard Deviations of Variables in Analysis
for the Three Language Groups

Space Topic

	Chinese (N= 73)		English (N= 89)		Spanish (N= 35)	
Variables	Mean	SD	Mean	SD	Mean	SD
Holistic Writing	2.78	1.05	5.60	.69	3.70	1.10
Holistic Reasoning	2.48	.79	4.47	.86	3.03	.61
Purves/Soter Scheme						
Content	14.18	3.12	22.30	3.00	17.53	3.27
Organization	6.40	1.55	9.92	1.28	7.65	1.53
Style and Tone	5.71	1.08	9.37	1.40	6.69	1.32
Moss Scheme						
Claims	4.79	1.93	6.88	2.23	5.52	1.46
Support	1.60	1.08	3.69	1.33	2.23	.97
Qualifications	.24	.34	.70	.64	.22	.32
Writer's Workbench Variables						
Number of suggestions	4.04	3.26	8.21	5.08	5.09	3.18
No. spelling errors	2.97	2.17	2.47	2.74	3.34	2.83
% vague words	8.24	2.92	5.59	1.83	6.56	1.69
Number to check	1.50	1.29	3.38	1.80	2.34	1.49
Avg. sentence length	18.66	7.61	21.90	4.81	23.96	7.27
% shorter sentences	23.28	16.37	28.13	11.40	34.23	15.09
% longer sentences	9.29	9.19	12.29	6.15	13.51	9.84
% "to be" verbs	77.18	12.33	84.22	10.40	80.97	11.59
% passive voice	6.11	6.81	15.26	8.76	11.83	8.65
% nominalizations	3.68	1.63	3.52	1.21	3.51	1.48
Avg. word length	4.49	2.71	4.90	2.26	4.61	2.73
% content words	54.60	3.56	56.11	3.02	52.65	4.47
Avg. length content words	5.88	4.02	6.48	3.23	6.18	4.39
% prepositions	9.26	2.28	11.99	1.59	11.72	2.60
% adjectives	14.53	2.85	16.20	2.84	13.54	3.47
% abstract	3.55	1.67	3.50	1.17	3.59	1.44

(GRE General Test Scores on Table 7)

Table 7

Means and Standard Deviations of Variables in Analysis for the Three Language Groups

Farming Topic

	Chinese		English		Spanish	
	(N= 73)		(N= 89)		(N= 35)	
Variables	Mean	SD	Mean	SD	Mean	SD
Holistic Writing	2.92	1.04	5.78	.52	3.61	1.10
Holistic Reasoning	2.49	.78	4.66	.84	2.81	.89
Purves/Soter Scheme						
Content	13.58	3.00	21.67	3.50	16.29	3.14
Organization	6.08	1.49	10.00	1.79	7.00	1.90
Style and Tone	5.97	1.16	9.35	1.28	6.99	1.34
Moss Scheme						
Graph Reading	3.51	1.30	4.02	1.16	3.30	1.77
Deductions	.61	.72	1.22	1.17	.64	.82
Claims	2.20	1.15	3.23	1.02	2.46	1.25
Justifications	1.14	.88	2.14	1.16	1.34	.97
Qualifications	.03	.10	.26	.30	.11	.18
Writer's Workbench Variables						
Number of suggestions	3.49	3.29	9.00	5.31	5.60	3.63
No. spelling errors	1.79	1.79	1.82	2.46	3.49	3.25
% vague words	5.59	3.27	4.82	2.03	5.76	2.67
Number to check	1.18	1.21	2.71	1.57	2.37	1.90
Avg. sentence length	19.68	6.89	22.33	6.53	24.93	6.24
% shorter sentences	26.10	14.68	27.91	10.86	32.23	14.04
% longer sentences	10.22	9.11	11.21	8.12	14.83	10.87
% "to be" verbs	71.83	19.91	79.64	16.38	69.54	20.17
% passive voice	10.10	12.36	13.74	9.05	8.00	7.06
% nominalizations	2.96	1.30	2.53	1.25	3.17	1.44
Avg. word length	4.55	2.60	4.78	2.23	4.51	2.50
% content words	57.99	3.87	58.34	3.53	54.37	3.50
Avg. length content words	5.63	7.38	6.06	6.58	5.82	3.37
% prepositions	12.22	3.77	14.13	2.47	13.77	3.61
% adjectives	17.17	4.07	18.46	3.05	16.39	3.49
% abstract	3.15	1.40	2.67	1.22	2.93	1.39
GRE General Test Scores						
Verbal						
Sentence completion	3.60	1.63	10.03	2.94	6.29	2.81
Discrete verbal	11.18	4.87	25.87	6.64	18.29	5.69
Reading comprehension	6.62	3.03	16.63	3.33	10.86	4.01
Analytical Reasoning						
Analytical reasoning	19.89	6.09	24.02	6.15	17.26	5.99
Logical reasoning	3.74	1.47	7.93	2.26	4.86	2.17

Table 8

Stepwise Regression Analyses Predicting Holistic Writing Scores from 21 Writer's Workbench Variables (before final reduction to 16 variables)

Space Topic

Total Sample (N=203)

Step	Writer's Workbench Independent Variables	R Squared	Adjusted R Squared	beta
1	Number of words	.55	.54	.74
2	Avg. word length	.66	.66	.37
3	Number of spelling errors	.69	.69	-.17
4	Percent passives	.71	.70	.16
5	Percent nominalizations	.71	.71	-.10
6	Kincaid readability	.72	.71	-.08
7	Percent simple sentences	.73	.72	-.10
8	Percent vagueness	.73	.72	-.09
9	Percent adjectives	.74	.73	.08

(F= 60.22 with 9, 193 df, p=.001)

Chinese Sample (N=73)

Step	Writer's Workbench Independent Variables	R Squared	Adjusted R Squared	beta
1	Number of words	.19	.18	.44
2	Percent abstract	.26	.24	-.25
3	Percent longer sentences	.32	.29	-.25
4	Percent vague words	.36	.32	-.21
5	Percent adjectives	.39	.35	.17
6	Avg. sentence length	.43	.37	-.22

(F= 8.17 with 9, 193 df, p=.001)

English Sample (N=89)

Step	Writer's Workbench Independent Variables	R Squared	Adjusted R Squared	beta
1	Number of words	.09	.08	.30
2	Percent adjectives	.14	.12	.22*

(F= 6.79 with 9, 193 df, p=.001)

Spanish Sample (N=35)

Step	Writer's Workbench Independent Variables	R Squared	Adjusted R Squared	beta
1	Percent passives	.28	.26	.53
2	Percent shorter sentences	.51	.48	.48

(F= 16.70 with 9, 193 df, p=.001)

*Asterisks indicate variables with negative beta weights that were positively correlated with the criterion

Table 9

Stepwise Regression Analyses Predicting Holistic Writing
Scores from 21 Writer's Workbench Variables
(before final reduction to 16 variables)

Farming Topic

Total Sample (N=203)

	Writer's Workbench		Adjusted	
Step	Independent Variables	R Squared	R Squared	beta
1	Number of words	.45	.45	.67
2	Average word length	.55	.54	.31
3	Percent prepositions	.58	.58	.19
4	Percent nominalizations	.61	.60	-.17
5	Number of suggestions	.62	.61	.13
6	Number of sentences	.64	.63	.23
7	Percent vague words	.64	.63	-.10
8	Av. length content words	.65	.64	.10
9	Number of spelling errors	.66	.64	-.09

(F= 40.94 with 9, 193 df, p=.001)

Chinese Sample (N=73)

	Writer's Workbench		Adjusted	
Step	Independent Variables	R Squared	R Squared	beta
1	Number of sentences	.11	.10	.34
2	Percent prepositions	.28	.26	.43
3	Av. length content words	.32	.29	.21
4	Number of spelling errors	.35	.31	-.19

(F= 9.27 with 9, 193 df, p=.001)

English Sample (N=89)

	Writer's Workbench		Adjusted	
Step	Independent Variables	R Squared	R Squared	beta
1	Number of words	.22	.22	.47
2	Percent shorter sentences	.31	.29	.29
3	Percent content words	.35	.33	.21
4	Percent "to be" verbs	.37	.34	.15

(F= 12.22 with 9, 193 df, p=.001)

Spanish Sample (N=35)

	Writer's Workbench		Adjusted	
Step	Independent Variables	R Squared	R Squared	beta
1	Number of sentences	.45	.43	.67
2	Number of suggestions	.51	.48	.25

(F= 16.42 with 9, 193 df, p=.001)

Table 10

Stepwise Regression Analyses Predicting GRE Analytical Reasoning Scores from All Variables (GRE General Verbal Part Scores and Writing Sample Scoring Schemes) for Total Sample (N=203)

Space Topic

Step	Writer's Workbench Independent Variables	R Squared	Adjusted R Squared	beta
1	GRE reading comprehension	.18	.18	.43
2	WWB percent passives	.21	.20	-.17*
3	WWB percent shorter sentences	.23	.22	-.14
4	Holistic reasoning	.24	.23	.18

(F= 15.90 with 4,198 df, p=.001)

Farming Topic

Step	Writer's Workbench Independent Variables	R Squared	Adjusted R Squared	beta
1	GRE reading comprehension	.18	.18	.43
2	WWB avg. sentence length	.22	.21	-.20
3	Moss claims	.24	.23	.17
4	Moss deductions	.27	.26	.18
5	WWB avg. length content words	.29	.27	-.13

(F= 15.98 with 4,198 df, p=.001)

Stepwise Regression Analyses Predicting GRE Logical Reasoning Scores from All Variables (GRE General Verbal Part Scores and Writing Sample Scoring Schemes) for Total Sample (N=203)

Space Topic

Step	Writer's Workbench Independent Variables	R Squared	Adjusted R Squared	beta
1	GRE reading comprehension	.60	.60	.77
2	GRE sentence completion	.65	.65	.40
3	Holistic writing	.66	.66	.13
4	Moss support	.67	.66	-.13*
5	Purves/Soter style and tone	.68	.67	.14
6	Moss qualifications	.68	.67	-.07*

(F= 69.43 with 4,198 df, p=.001)

Farming Topic

Step	Writer's Workbench Independent Variables	R Squared	Adjusted R Squared	beta
1	GRE reading comprehension	.60	.60	.77
2	GRE sentence completion	.65	.65	.40
3	Holistic writing	.66	.66	.15
4	WWB percent "to be" verbs	.67	.66	.07

(F= 99.98 with 4,198 df, p=.001)

*Asterisks indicate variables with negative beta weights that were positively correlated with the criterion

Table 11

Stepwise Regression Analyses Predicting GRE Analytical Reasoning and Logical Reasoning Scores from All Variables (GRE General Verbal Part Scores and Writing Sample Scoring Schemes) by Language Groups

Chinese Sample (N=73)

Analytical Reasoning as Dependent Variable

Space Topic

Step	Writer's Workbench Independent Variables	R Squared	Adjusted R Squared	beta
1	WWB percent passive voice	.05	.04	-.23
2	Holistic reasoning	.09	.06	.20
3	Moss support	.14	.10	-.32
4	WWB percent short sentences	.17	.12	-.19

(F= 3.50 with 4, 67 df, p=.001)

Farming Topic

Step	Writer's Workbench Independent Variables	R Squared	Adjusted R Squared	beta
1	Moss claims	.12	.11	.35
2	WWB percent short sentences	.18	.15	-.24
3	WWB percent adjectives	.24	.20	-.24
4	WWB percent passive voice	.29	.24	-.23
5	Moss graph reading skills	.33	.28	.21

(F= 6.50 with 5, 66 df, p=.001)

Logical Reasoning as Dependent Variable

Space Topic

Step	Writer's Workbench Independent Variables	R Squared	Adjusted R Squared	beta
1	GRE sentence completion	.16	.15	.41
2	WWB avg. word length	.22	.20	-.24
3	GRE reading comprehension	.29	.26	.27
4	Holistic writing	.33	.29	.21
5	WWB percent abstract	.36	.32	.20
6	Moss support	.40	.34	-.24*
7	WWB percent longer sentences	.43	.37	.18

(F= 6.84 with 7, 64 df, p=.001)

Farming Topic

Step	Writer's Workbench Independent Variables	R Squared	Adjusted R Squared	beta
1	GRE sentence completion	.16	.15	.41
2	GRE reading comprehension	.21	.19	.23
3	WWB avg. sentence length	.25	.22	-.21
4	Purves/Soter content/thinking	.29	.25	-.21
5	WWB avg. word length	.34	.29	-.22
6	Moss graph reading skills	.37	.31	.18
7	WWB percent abstract	.40	.33	.17

(F= 5.99 with 7, 64 df, p=.001)

*Asterisks indicate variables with negative beta weights that were positively correlated with the criterion.

Table 12

Stepwise Regression Analyses Predicting GRE Analytical Reasoning and Logical Reasoning Scores from All Variables (GRE General Verbal Part Scores and Writing Sample Scoring Schemes) by Language Groups

English Sample (N= 89)

Analytical Reasoning as Dependent Variable

Space Topic

	Writer's Workbench		Adjusted	
Step	Independent Variables	R Squared	R Squared	beta
1	GRE sentence completion	.14	.13	.38
2	Moss qualifications	.20	.18	-.24
3	WWB percent longer sentences	.25	.23	-.23
4	WWB percent nominalizations	.30	.27	-.23

(F= 9.38 with 4, 84 df, p=.001)

Farming Topic

	Writer's Workbench		Adjusted	
Step	Independent Variables	R Squared	R Squared	beta
1	GRE sentence completion	.14	.13	.38
2	WWB avg. length content words	.24	.22	-.30
3	WWB number to check	.27	.25	.19

(F= 10.99 with 3, 85 df, p=.001)

Logical Reasoning as Dependent Variable

Space Topic

	Writer's Workbench		Adjusted	
Step	Independent Variables	R Squared	R Squared	beta
1	GRE sentence completion	.33	.33	.58
2	GRE reading comprehension	.39	.38	.33
3	Moss qualifications	.42	.40	-.17

(F= 21.41 with 3, 85 df, p=.001)

Farming Topic

	Writer's Workbench		Adjusted	
Step	Independent Variables	R Squared	R Squared	beta
1	GRE sentence completion	.33	.33	.58
2	GRE reading comprehension	.39	.38	.33
3	Moss graph reading	.42	.40	-.14
4	Moss qualifications	.44	.41	.16
5	WWB percent content words	.46	.43	.17
6	Moss justifications	.48	.44	-.14

(F= 13.15 with 6, 82 df, p=.001)

Table 13

Stepwise Regression Analyses Predicting GRE Analytical Reasoning and Logical Reasoning Scores from All Variables (GRE General Verbal Part Scores and Writing Sample Scoring Schemes) by Language Groups

Spanish Sample (N=35)

Analytical Reasoning as Dependent Variable

Space Topic

Step	Writer's Workbench Independent Variables	R Squared	Adjusted R Squared	beta
1	GRE reading comprehension	.44	.42	.66
2	WWB percent nominalizations	.53	.50	-.30
3	WWB percent adjectives	.61	.57	.28

(F= 16.00 with 3, 31 df, p=.001)

Farming Topic

Step	Writer's Workbench Independent Variables	R Squared	Adjusted R Squared	beta
1	GRE reading comprehension	.44	.42	.66
2	Moss deductions	.55	.52	.33

(F= 19.29 with 2, 32 df, p=.001)

Logical Reasoning as Dependent Variable

Space Topic

Step	Writer's Workbench Independent Variables	R Squared	Adjusted R Squared	beta
1	GRE reading comprehension	.35	.33	.59
2	Holistic writing	.44	.40	.32
3	WWB percent vague words	.56	.52	.37
4	Moss support	.61	.56	-.23
5	Purves/Soter style and tone	.69	.64	.31
6	WWB percent nominalizations	.73	.67	.19
7	WWB avg. length content words	.77	.71	-.28

(F= 12.76 with 7, 27 df, p=.001)

Farming Topic

Step	Writer's Workbench Independent Variables	R Squared	Adjusted R Squared	beta
1	GRE reading comprehension	.35	.33	.59
2	Holistic writing	.49	.46	.40
3	WWB percent content words	.55	.50	-.25
4	Holistic reasoning	.59	.54	-.26*
5	Purves/Soter style and tone	.66	.60	.30
6	GRE sentence completion	.71	.65	.29
7	WWB avg. sentence length	.75	.69	.22
8	WWB percent abstract	.79	.72	.22
9	Moss justifications	.81	.74	.22

(F= 11.93 with 9, 25 df, p=.001)

*Asterisks indicate variables with negative beta weights that were positively correlated with the criterion

Appendix A

Space and Farming Topics

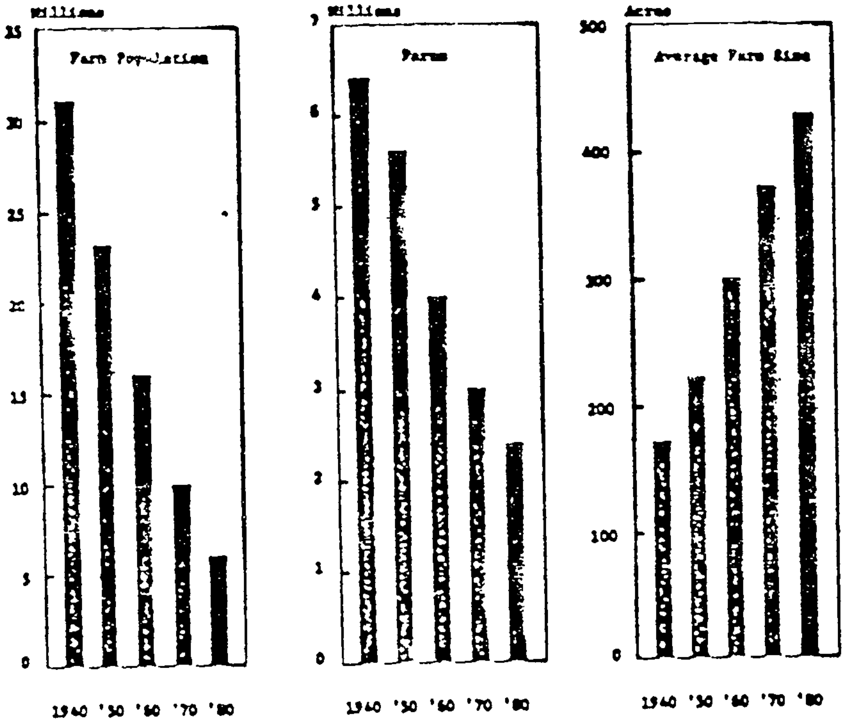
TIME - 30 MINUTES

Some people say that exploration of outer space has many advantages; other people feel that it is a waste of money and other resources. Write a brief essay in which you discuss each of these positions. Give one or two advantages and disadvantages of space exploration, and explain which position you support.

THIS SPACE MAY BE USED FOR NOTES.

TIME - 30 MINUTES

CHANGES IN FARMING IN THE U.S.: 1940 - 1980



Suppose that you are writing a report in which you must interpret the three graphs shown above. Write the section of that report in which you discuss how the graphs are related to each other and explain the conclusions you have reached from the information in the graphs. Be sure the graphs support your conclusions.

THIS SPACE MAY BE USED FOR NOTES.

Appendix B

Purves/Soter Scoring Scheme

A. Content/Thinking

This dimension concerns what is written and the way it reflects the writer's manipulation of the subject. It is divided into a number of specific aspects that should be scored separately.

1. Adequacy of information presented refers to the degree to which all of the relevant information from the stimulus is contained in the text.

2. Richness of additional information refers to the use of additional information to that which is in the stimulus (e.g., information drawn from a variety of sources such as reading or general knowledge) and may be seen as the amount of relevant allusion.

3. Relationships drawn refers to the degree to which the text shows that connections have been made between or among the various items of information and the validity and/or complexity of the relationships.

4. Inferences made refers to the number and depth of interpretations (causal, resultant, comparative, contrastive, extrapolative) that the writer makes beyond the information in the stimulus and or from the outside.

5. Synthesis refers to the degree to which the writer appears to draw together the information, relationships, and inferences into a single or complex generalization.

6. Evaluation refers to the degree to which the writer appears to make judgments as to the relative merit of particular relationships, inferences, or syntheses and the degree to which applicable criteria are used.

7. Consideration of alternatives refers to the extent to which the writer appears to admit the possibility of alternative or counter-arguments (as in Space) or interpretations (as in Farms) and either accepts them as admissible or rebuts them.

B. Organization

This dimension concerns the structure of the written text both as a whole text and in its various parts.

8. Framing refers to the degree to which the writer presents a context for the content in such a manner that there is an apparent beginning, middle, and end. There need not be a formal introduction or conclusion, but there should be a clear sense that the composition begins and ends appropriately.

9. Grouping refers to the degree to which the writer joins (in paragraphs or through some other means) the various pieces of information, relationships, or inferences). Inadequate combining would be represented by a list in which there is no discernible pattern or system of combining bits of information.

10. Unity refers to the degree to which the writer appears to have both grouped appropriate information, relationships, and inferences and excluded extraneous matter or thoughts. The reader may see this in units such as paragraphs and/or in the composition as a whole.

C. Style/Tone

This dimension refers to the manner in which the composition is presented, and particularly to the degree to which that manner matches the conventions of academic discourse in the United States.

11. Objectivity refers to the use of impersonal and detached language as opposed to personal and emotional language. An objective composition may use the first person singular occasionally, but usually as a qualifier.

12. Tentativeness refers to the use of semantic hedges and qualifiers that are often considered appropriate to academic writing. In the responses to the particular stimuli of these assignments, the better composition will probably not use extreme or dogmatic language.

13. Metalanguage refers to the use of markers (e.g., "however") to indicate the relationship between propositions and paragraphs. These markers would include connectives and markers that illustrate such relationships as cause, effect, comparison, contrast classification, definition, or hypothesis. The markers would signal the sequence and structure of the composition.

Appendix C

Moss Scoring Scheme

Space Topic

ID ______ Reader ________

(+ = existence; blank = nonexistence)

Advantages

___ Brief descriptor________________________________

___ clear; nonambiguous

___ accurate fact or reasonable inference; relevant

___ elaborated with examples (or a more specific extension of the advantage without justification/explanation)

___ justified/explained (e.g., "this is an advantage because...")

___ qualified/rebutted (i.e., shown not to be an advantage)

___ builds on previous advantage (to justify/support current advantage)

___ integrates as evidence at least two previous advantages

[continues for all descriptors]

[on subsequent pages]

Disadvantages

___ Brief descriptor________________________________

___ clear; nonambiguous

___ accurate fact or reasonable inference; relevant

___ elaborated with examples (or a more specific extension of the disadvantage without justification/explanation)

___ justified/explained (e.g., "this is a disadvantage because...")

___ qualified/rebutted (i.e., shown not to be a disadvantage)

___ builds on previous disadvantage (to justify/support current disadvantage)

___ integrates as evidence at least two previous disadvantages

---

Opinion/Point of View

___ Brief descriptor________________________________

___ explicit summary statement of position made

___ position clear, nonambiguous (scoring can be based on an explicit statement of position or on a position implied in presentation of advantages and disadvantages)

___ criteria explicit (advantages/disadvantages explicitly connected to opinion)

___ criteria implicit (advantages/disadvantages implicitly connected to opinion)

___ advantages/disadvantages rebutted or qualified in ways consistent with opinion (explicit recognition of competing priorities)

Abridged Version of Moss Scoring Scheme

Farming Topic

ID ______ Reader ________

(+ = existence; blank = nonexistence)

Descriptions of/Deductions from Graphs

___ accurate description of graph 1 (farm population decreased)

___ specific description of graph 1 (at least three pieces of information on two data points)

___ accurate description of graph 2 (number of farms decreased)

___ specific description of graph 2 (at least three pieces of information on two data points)

___ accurate description of graph 3 (size of farms increased)

___ specific description of graph 3 (at least three pieces of information on two data points)

___ relationship between graphs 1 and 2 explicit

___ new variable or quantity deduced from graphs 1 and 2

___ relationship between graphs 1 and 3 explicit

___ new variable or quantity deduced from graphs 1 and 3

___ relationship between graphs 2 and 3 explicit

___ new variable or quantity deduced from graphs 2 and 3

___ relationship between graphs 1, 2, and 3 explicit

___ new variable or quantity deduced from graphs 1, 2, and 3

Conclusion(s)

Conclusions include statements that draw inferences from the graphs (e.g., of cause, consequence, prediction) that may require additional information or justification for support.

___ Brief descriptor ____

____

- ___ clear; nonambiguous
- ___ reasonable inference
- ___ evidence provided or requested/justification given (in addition to data from the graphs)
- ___ elaborated with examples (or a more specific conclusion/extension of the conclusion without justification)
- ___ obvious connection made between data from graphs and conclusion
- ___ qualified/pcssible rebuttal mentioned (e.g., labeled as "likely" or "possible" conclusion, rather than "the" conclusion; or as a likely conclusion "unless" certain conditions change)
- ___ builds on previous conclusion(s) (to justify/support current conclusion)
- ___ integrates as evidence at least two previous conclusions

[continues for however many conclusions are presented]

Appendix D
Reid Holistic Reasoning Schemes

Space Topic

General Considerations

Content

Quality of analysis
of detail
of logic
of communication of ideas
of focus

Length: breadth, depth, and appropriateness of detail; limits and focuses topic

Process/Organization

May not fulfill expectations of the U.S. academic audience, but must have the overall impact of completeness and the sequencing of ideas in a logical manner

Describes/states the problem: introduction (may be implied)
Provides statement of focus (topic sentence): choosing an alternative
Develops ideas by specific detail, examples, explanation
Observes/explains the choices
Analyzes the value/benefits; provides supporting analysis with detailed material
Concludes with summary, solution, question, prediction; based on earlier arguments and/or examples

Space

Scoring

Pluses that contribute to the holistic evaluation

Unusually tight (vs. unusually loose)
Unusually rich (vs. paucity of development or examples); may develop the argument with examples
Originality within parameters of task (vs. regurgitating topic)

High-range Papers (5-6 on a 1-6 scale)

Weighted persuasive presentation (or perhaps equally presented both sides)
Adequate, appropriate support (facts, examples, physical description, and/or personal experience)
High degree of overall fluency (?)
Must include one or more of the +'s

Note: Some high papers don't take a stand on the alternatives; sometimes they present a balanced explanation of both sides, then leave the choice to the reader (they don't presume to choose).

Middle-range Papers (4-5 on a 1-6 scale)

May repeat the assignment as the statement of introduction
Has a sense of direction and some detail, but focus or support is thin or nonexistent
May state opinions without support
Overall fluency flawed (?)
For a 4: basic elements are there without +'s
For a 3: some basic elements are there without +'s

Low-range Papers (1-2 on a 1-6 scale)

Organization and focus flawed seriously or nonexistent
Little or no knowledge of what compare/contrast demands
Little or no ability to assume and produce the task
Overall fluency extremely limited or nonexistent (?)
Serious lack of basic elements

D.3

Holistic Reasoning Scoring Scheme
Farming Topic

General Considerations

Content

Quality of analysis: detail, logic, communication of ideas, focus

Length: breadth, depth, and appropriateness of detail; limits and focuses topic

Process/Organization: May not fulfill expectations of the U.S. academic audience, but must have the overall impact of completeness and the sequencing of ideas in a logical manner

States the problem or issue (may be implied)

Describes the chart(s): overall description

Discusses the chart(s): more specific details of description that links the charts

Analyzes the chart(s): integrates the information, provides causal connections

Concludes: possible solution, summary, provocative questions, predictions; based on earlier or following arguments and/or examples

Pluses (+) that contribute to the holistic evaluation:

Unusually tight (vs. unusually loose)

Unusually rich (vs. paucity of development or examples); may develop the argument with examples

Originality within parameters of task (vs. restating topic)

High-range Papers (5-6 on a 1-6 scale)

Clear organization; may make direct reference to the chart(s) and/or assignment

Careful analysis; adequate support for analysis

Overall fluency high--focus on presentation of ideas vs. vocabulary

Must include one of more of the +'s

Note: Some high papers don't analyze (don't presume to do so with so little information); some high papers don't describe in great detail the charts--they get right to the business of analysis (implying a degree of audience background).

Middle-range Papers (4-5 on a 1-6 scale)

Clear, detailed description; perhaps limited discussion

Some analysis, but little or no support for the analysis

Overall fluency flawed

A score of 4: basic elements are there without +'s

A score of 3: some basic elements are there without +'s

Note: At least 2 of these elements are usually present in middle-range papers, often more

Low-range Papers (1-2 on a 1-6 scale)

Description but little or no discussion or analysis

Organization is not evident; simple chronology at best

Little or no detail; serious lack of elements

Summary follows immediately, if at all

Overall fluency is very limited or non-existent

Serious mechanical errors interfere with basic communication of ideas

297977