Received: December 14, 2011 *Accepted: February 7, 2012*

A Comparison of the Performance of Analytic vs. Holistic Scoring Rubrics to Assess L2 Writing

Cynthia S. Wiseman[1]

Abstract

This study compared the performance of a holistic and an analytic scoring rubric to assess ESL writing for placement and diagnostic purposes in a community college basic skills program. The study used Rasch many-faceted measurement to investigate the performance of both rubrics in scoring second language (L2) writing samples from a departmental final examination. Rasch analyses were used to determine whether the rubrics successfully separated examinees along a continuum of L2 writing proficiency. The study also investigated whether each category in the two six-point rubrics were useful. Both scales appeared to be measuring a single latent trait of writing ability. Raters hardly used the lower category of the holistic rubric, suggesting that it might be collapsed to create a five-point scale. The six-point scale of the analytic rubric, on the other hand, separated examinees across a wide range of strata of L2 writing ability and might therefore be the better instrument in assessment for diagnostic and placement purposes.

Keywords: *L2 writing assessment, Analytic scoring, Holistic scoring, Rubrics, Rasch, MFRM*

1. Introduction

A direct writing assessment is a performance-based test that involves multiple components in any assessment situation, including the writer/examinee, the task, the raters and the rating procedure (Hamp-Lyons, 2003) with the scoring rubric being a key subcomponent in the assessment of direct writing. Holistic scoring is a global approach to the text that reflects the idea that writing is a single entity which is best captured by a single scale that integrates the inherent qualities of the writing, and that this quality can be recognized only by carefully selected and experienced readers using their skilled impressions, rather than by any objectifiable means (White, 1985; Weigle, 2002; Hyland, 2002). Some have argued that holistic scoring focuses on what the writer does well rather than on the writer's specific areas of weakness,

[1] *City University of New York.. Email: cwiseman@bmcc.cuny.edu , engspeak@aol.com*

which is of more importance for decisions concerning promotion or placement (Charney, 1984; Cumming, 1990; Hamp-Lyons, 1990; Reid, 1993, Cohen & Manion, 1994; Elbow, 1999). White (1985) argued that holistic scoring focuses the attention on the strengths of the writing rather than the deficiencies, and he asserted that holistic scoring ‘reinforces the vision of reading and writing as intensely individual activities involving the full self” and that any other approach is ‘reductive’ (p. 409). From this perspective, a holistic scoring method may often be the choice of not only writing faculty but also program administrators, who may often choose holistic scoring rubrics for L2 writing assessment for practical reasons; that is, it is more economical to assign one score to an essay by reading it once; indeed, holistic scoring rubrics are widely used for large-scale exams (Godshalk, Swineford & Coffman, 1966; Alloway, 1978; Powills, Bowers & Conlan, 1979).

While a holistic scoring method could serve the economic interests of a program, a single score based on a holistic reading of the essay may not serve the best interests of L2 writers/examinees. Holistic scoring does not allow raters to distinguish between various aspects of writing such as control of syntax, depth of vocabulary mastery, and organizational control. Yet, these variables may influence scores. Indeed, Barkaoui (2010) found that individual, textual and contextual factors in the rating context introduced variability in holistic scores of L2 writing samples.

For second language learners, this is problematic since different aspects of writing ability develop at different rates for different writers. Some writers may be strong in expressing content and organization but limited in grammatical accuracy, while others may have excellent language control at the sentence level but are unable to organize their writing. Some learners may not perform the same in each of the components of the writing skill (Kroll 1990), which makes more qualitative evaluation procedures such as lexical, syntactic, discourse and rhetorical features necessary (Reid 1993).

As an alternative, analytical scoring methods, in which raters make judgments about nominated features or writing skills, involve the separation of the various features of a composition into components for scoring purposes. Writing samples are thus rated on an analytic rubric that includes several domains representing the construct of writing (Weigle, 2002). Analytic scoring rubrics, thus, provide more information about a test taker’s performance than the single score of a holistic rating and permit a profile of the areas of language ability that are rated. For that reason, analytic scoring methods are often chosen for placement and diagnostic purposes (Jacobs, Zingraf, Warmuth, Hartfiel & Hughey, 1981; Perkins, 1983; Hamp-Lyons, 1991). Indeed, comparing analytic scales, Knoch (2009) found that rater reliability was substantially higher and that raters were able to better distinguish between different aspects of writing when the more detailed descriptors of the analytic scale were used. Rater feedback also showed a preference for the more detailed scale. While there appear to be advantages to analytic scoring for second language writing ability, e.g., a more individualized profile of the L2 writer, there is often resistance given the increased cost in time and money.

While many examples of rating scales for second language exist for writing proficiency (Shohamy, 1995), it has been noted that rating scales used in subjective scoring present major problems of reliability and validity (Bachman & Savignon, 1986; Fulcher, 1987; Matthews, 1990). Smith, Winters-Edys, Quellmalz, & Baker (1980) in a comparison of alternative methods for placing post-secondary students into freshman English or remedial writing examined the comparability of scores obtained from three scoring methods and found relationships among the scores from the three methods to be low. In a study by Huang (2008) investigating the holistic ratings of ESL and NE students' writing, he found differences in terms of consistency and precision. These findings strongly suggested the need for scrutiny of the reliability and validity of placement standards, scoring criteria and the emphasis of each on essay features in high-stakes assessment contexts.

Routinely, test developers of standardized writing examinations, such as IELTS, have dedicated themselves to reliability and validation studies of scales used in scoring writing and speaking (Shaw, 2002). Like for any standardized high stakes assessment, it is incumbent upon a college ESL program that administers a high stakes test to investigate the performance of that assessment.

While the choice of one type of rubric or the other is often determined by considerations of practicality, this study examine the performance of a holistic vs. an analytic scoring rubric in the assessment of L2 writing ability for placement and diagnostic purposes. Would the holistic rubric used to score a set of writing samples adequately separate examinees across strata of proficiency? Would the analytic rubric comparably separate the same sample of examinees in terms of writing proficiency? Which of those two types of scales would better discriminate examinees by proficiency level?

2. Background Context and Rationale

This study was conducted in a large urban community college serving approximately 19,000 students in degree programs with 6,000 more in continuing education programs. The college served an international population with students from over 100 countries. Asian, Hispanic and Black racial/ethnic groups, according to student self-descriptions, comprised greater than 85% of the student population. Many of these students spoke English as a second language.

Those non-native English-speaking students who did not pass a basic skills writing exam were required to take ESL courses. Placement in ESL was routinely determined by a third reading of a placement exam by a faculty reader using the department holistic rubric. Based on that faculty's assignment of a single holistic score to the writing sample, the student was then placed in a particular level of ESL intensive writing or sent to the English remedial skills department if, based on the writing sample, it was determined that the writer was a native speaker of English.

After the initial assignment of students to their respective levels, an in-class diagnostic essay was administered the first day of class to identify students who had been misclassified. While this process served to identify many misclassified students, it was not without its problems. In a typical semester out of approximately 600 students placed in ESL, 110-120 were misplaced, a little more than one-sixth of the incoming students. In effect, in a class of 25, there might be four or five transfers. The number of misplaced students suggested the need to reexamine the scoring procedures for making those decisions, and in particular, to explore whether the use of an analytic scoring procedure would provide additional information that could improve accuracy in placement and promotion.

2.1 L2 Writing Proficiency

In the creation of an analytic rubric for the evaluation of L2 writing, the first step was to define the nature of L2 writing ability. L2 writing ability was defined within the framework of Bachman and Palmer's (1996) model of communicative language ability (CLA) as a specific combination of language ability and task characteristics, that is, the language ability required in the contextualized performance of a task, which, in this case, was writing a composition. Writing an essay requires topical knowledge, which Bachman and Palmer (1996) define as "knowledge schemata" or "real-world knowledge," as well as strategic competence, or planning and executive strategies to develop the topic of the composition using the L2. It also requires textual, or rhetorical, knowledge to organize the supporting propositions, as well as grammatical knowledge of the L2, demonstrated as grammatical control in writing. The task of writing a composition also requires pragmatic knowledge and sociolinguistic competence that allow the L2 writer to use the vocabulary and register appropriate to the audience. L2 writing ability was thus defined in terms of topical knowledge and strategic competence, or content development; textual knowledge, or cohesion and rhetorical organization; sociolinguistic competence, or knowledge of vocabulary and register appropriate to academic writing; and grammatical knowledge.

2.2 Multi-Faceted Rasch Measurement Model

The many-faceted Rasch model was used to examine the use of holistic vs. analytic scoring procedures in this multi-faceted assessment. This model makes possible the analysis of data from assessments with multiple facets, such as, in this case, examinees, raters, rubrics and essay prompts. The model views each score obtained by an examinee on an L2 writing assessment as the result of the interaction between that particular examinee's ability, the severity of the reader who awarded the score, and the difficulty of the essay prompt. The ability of each examinee is thus calculated based on the likelihood of receiving a particular score on a given prompt, taking into account that prompt's difficulty and also the severity of the rater assigning the score. Similarly, the severity of each rater could be understood as the probability of the rater awarding a given score to an examinee of a particular ability who responded to a prompt of a certain difficulty or the difficulty of the essay prompts could be expressed as a function of the likelihood of an examinee of a particular ability receiving a certain score (or better) on that essay prompt, from a reader of a given severity (McNamara, 1996).

3. Research Questions

Using the multi-faceted Rasch measurement model, several questions were asked to examine the performance of the holistic and analytic rubrics in this L2 writing assessment and to determine whether an analytic scoring rubric would make a significantly greater difference in the accuracy of placement and diagnostic decisions to reduce the number ESL writers who were misclassified each semester and alleviate the resulting problems when students are reclassified:

3.1 Examinees/Students

1. How much variability was there across student levels of proficiency? To what extent were examinees separated into distinct strata of proficiency? That is, how many different levels of examinee L2 proficiency were there?

3.2 Rating Scale

1. Did the analytic and holistic rating scales function properly as six-point scales? Were the scale categories clearly distinguishable, i.e., "most probable" over clearly defined intervals?

2. Were the 6 categories of both the analytic and holistic rating scales appropriately ordered?

3. Which, if either, of these two scales – the analytic or the holistic rubrics – would better separate the examinees across proficiency levels?

4. Method

4.1 Examinees

Test-takers (N=60) were matriculated non-native speakers of English enrolled in Developmental English (ESL) Intensive writing classes in a community college in a large urban setting. Sixty writing samples were randomly selected from a single administration of a departmental final examination administered among 3 levels of ESL classes: ESL 054, ESL 062, and ESL 094. The samples were taken from 8 different classes.

4.2 Raters

Raters of the writing samples included 5 experienced instructors of writing, all experienced and seasoned writing teachers with 15 to 30 years' teaching experience and considerable experience scoring exams. For more than 18 years, 3 of the 5 raters had previously participated in numerous departmental training sessions using the holistic scale at least twice each semester. A fourth rater was a newly hired junior faculty member who had used the holistic scale for only two semesters. The fifth rater was also a newly hired full-time faculty member with 7 years' service as an adjunct instructor of writing, but little experience in scoring writing samples with the holistic rubric. They all participated in the training and norming session introducing the analytic rubric.

The departmental final exam, a timed impromptu writing test of second language writing ability, was administered on the final day of the semester by all classroom teachers. Each examinee chose one of two prompts, which were retired items of the Writing Assessment Test (WAT), a basic skills proficiency examination used by the university to assess basic writing proficiency. Examinees had one hour to write either a persuasive or narrative essay. Each writing sample was then scored by two readers and assigned a composite score between two and twelve to determine promotion. Sixty writing samples were randomly selected from the approximately 450 exams.

4.3 Rating Rubrics

4.3.1 Holistic Rubric

The holistic rubric was a six-point holistic scale that provided a general narrative description of a typical paper for each score point. Performance criteria included organization (logical structure), development of content, vocabulary, use of rhetorical devices, sentence variety, language (e.g., agreement and grammatical inflection), punctuation, and paper length. The performance criteria, however, were not uniform across scale points, e.g., paper length was a performance criterion for a two-point paper but not a criterion for a six-point paper. (See Appendix A.)

4.3.2 Analytic Rubric

The analytic rubric was a six-point scale with five domains that represented the construct of second language writing as determined by a lengthy content-validation process that included research in existing writing rubrics, student writing samples, faculty input, alignment of domains with curricula, course objectives, and the current holistic rubric, as well as rater input. The newly constructed analytic rubric included five subdomains: Task fulfillment, topic development, organization, register and vocabulary, and language control. To ensure that the scale points were mutually exclusive, the performance criteria for each domain attempted to reflect differences in proficiency levels. (See Appendix B).

At the first norming session each rater was given a copy of the holistic rubric and a packet of sixty writing samples. At the second norming session, which took place one week later, each rater received a copy of the analytic rubric and the same set of sixty writing samples. The writing samples were numbered and any identifying information was deleted to protect the anonymity of the examinees.

4.4 Rating procedures

In both the holistic and analytic norming sessions, the five raters met to review the rubric. The performance criteria for each category were read aloud and discussed. Each rater then read and scored two writing samples. When scoring holistically, raters assigned a single score to each writing sample. When scoring analytically, raters were instructed to first read the compositions quickly to give a score for task fulfillment (overall impression), and then assign the two essays a score from 1-6 (1=low, 6=high) in each of the five domains and then to add the scores for a total score. The scores were then compared, and those who scored higher or lower than the norm explained the reasons for their score. Adjustments were made for scores above or below the consensus of the group. Once the five raters were adequately normed, the raters worked together in the group to score half of the sixty writing samples together. The raters scored the remaining essays at home. The scored writing samples and rating sheets were returned to the researcher within the week.

4.5 Statistical Procedures

First, descriptive statistics were computed to investigate whether scores assigned via both the holistic and analytic rubric were normally distributed. Internal consistency reliability (alpha) was also computed to examine how the five domains of the analytic rubric performed.

The main effects for the facets of examinee, rater, and prompt were examined using the FACETS (Linacre, 2005) computer program. Fit statistics were examined to identify any unusual rating patterns in light of the expected patterns predicted by the model. Rating scale functionality was also investigated by examining the average examinee proficiency measures.

The FACETS computer program is based on a many-faceted version of the Rasch measurement model for ordered response categories (Linacre, 1989), which is a generalization of the Rasch family of measurement models (Andrich, 1988; Rasch, 1980, Wright & Masters, 1982).

The Partial Credit form of the many-faceted Rasch model we utilized to analyze the analytic rubrics takes the form:

$$\ln\left(\frac{P_{njik}}{P_{njik-1}}\right) = \beta_n - \lambda_j - \delta_i - \tau_{ik}'$$

Where:

P_{njik} = the probability of examinee n being awarded a score of k when rated by reader j on essay prompt i

P_{njik-1} = the probability of examinee n being awarded a score of k-1 when rated by reader j on essay prompt i

β_n = the proficiency of examinee n

λ_j = the severity of reader j

δ_i = the difficulty of essay prompt i

τ_{ik} = the difficulty of achieving a particular score (k) averaged across all readers for each essay prompt separately

To analyze the holistic rubric, we used the Rating Scale form of the many-faceted Rasch model, which takes the form:

$$\ln\left(\frac{P_{njk}}{P_{njk-1}}\right)=\beta_n-\lambda_j-\tau_k$$

Where:

P_{njik} = the probability of examinee n being awarded a score of k when rated by reader j on essay prompt i

P_{njik-1} = the probability of examinee n being awarded a score of k-1 when rated by reader j on essay prompt i

β_n = the proficiency of examinee n

λ_j = the severity of reader j

τ_k = the relative probability of rating in category k as opposed to category k-1 for the scale when $\tau_1 = 0$.

The many-faceted Rasch model allows the researcher to establish a statistical framework that 1) summarizes overall rating patterns in terms of group-level main effects for each one of the facets, and 2) quantifies individual-level effects of the various components within each facet thus providing diagnostic information about how each individual examinee, rater, essay prompt and rubrics are performing (Engelhard &

Myford, 2003). For each element of a given facet, the FACETS computer program provides an estimate of a measure (in logits), a standard error (information concerning the precision of the measure), and fit statistics (information about how well the data fit the expectations of the measurement model). To examine how the rubrics performed, we have focused on two of the facets, namely, examinee writing ability and scoring rubric.

5. Results

5.1 Descriptive Statistics

The descriptive statistics for the scores assigned to the sixty writing samples using both the analytic and holistic rubric are presented in Table 1. For the holistic rubric, the mean was 30.5 with a standard deviation of 2.25. For the analytic rubric, the means ranged from 3.41 to 3.60 and the standard deviation from .95 to 1.14. All values for skewness and kurtosis for both rubrics were within accepted limits of ±2, indicating that scores assigned using each rubric were normally distributed.

Table 1. Descriptive Statistics for Scoring Rubrics

	Holistic Rubric	Analytic Rubric				
		Task Fulfillment	Topic Development	Organization	Vocabulary & Register	Linguistic Control
Mean	30.50	3.600	3.410	3.540	3.520	3.550
SD	2.25	1.140	1.120	1.030	.950	1.040
Skewness	.00	.068	.403	.314	.258	.206
Kurtosis	-1.20	-.318	-.324	-.192	-.305	-.133

The reliability estimate for internal consistency for both the holistic rubric (.81) and for the analytic rubric (.93) was quite high, which suggests that both rubrics are measuring a single construct.

5.2 The FACETS Analysis

The data analyses were designed around the research questions listed above. This discussion of the research findings in the FACETS analysis focuses on the main effects of examinee ability and scoring rubric in this L2 writing assessment.

To begin, the variable maps for both the analytic and holistic rubrics shown in Figures 1 and 2 provide a unified synopsis of the findings for all the facets of the analysis.

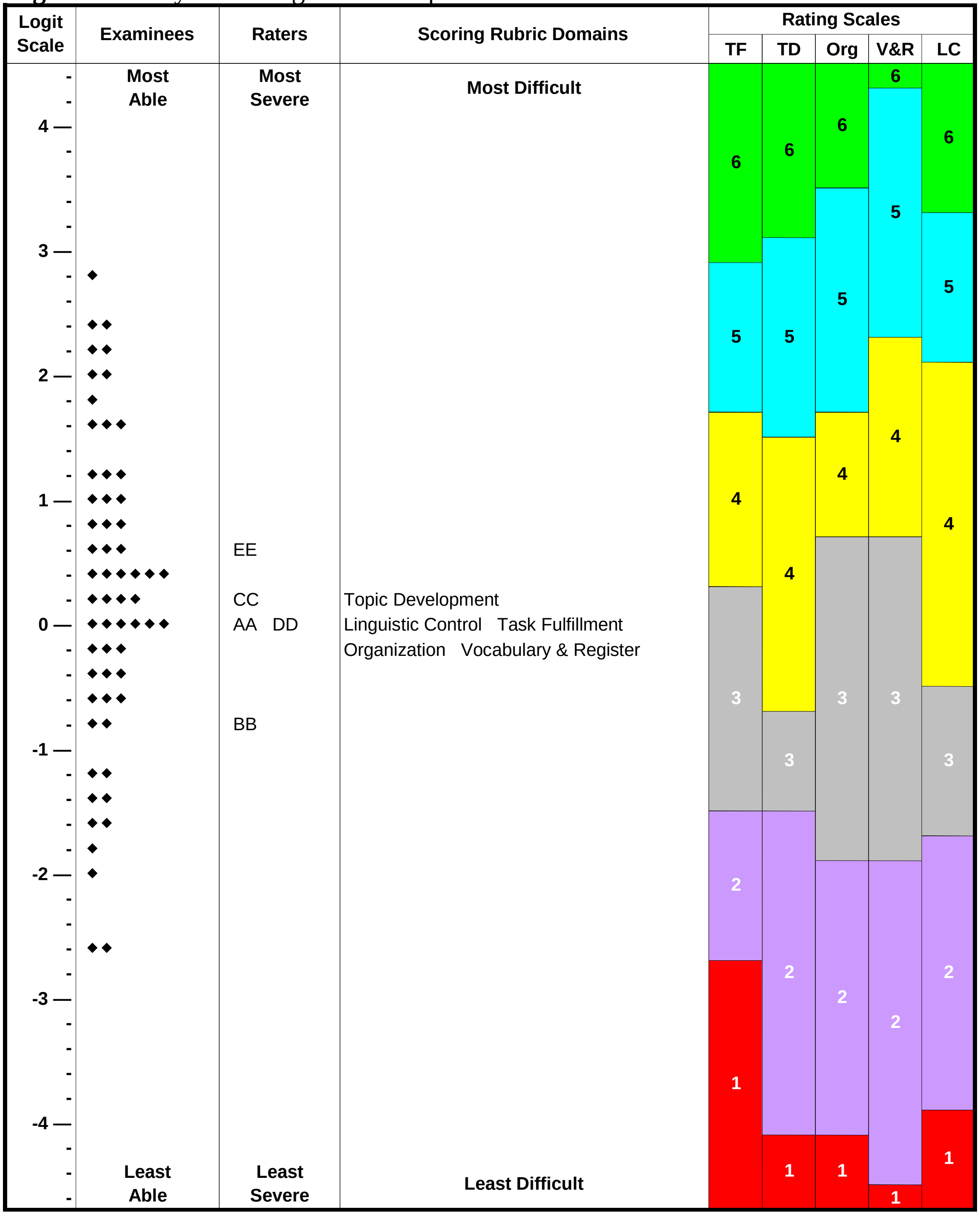
All facets of the assessment were all calibrated on the same scale, in particular, the facets of examinee ability and the performance of the rating scale for each scoring rubric. The unit of measurement on this scale is the “logit” which, as shown in equations (1) and (2), is obtained by a simple logarithmic transformation of the odds of receiving a particular score. When the data fit the model, the logit defines an equal-interval scale, which serves as a common frame of reference for all the facets of the analysis, thus facilitating comparisons within and between facets. The logit scale is displayed in the first column of the variable map.

The second column of the map displays the estimates of examinee proficiency on the respective domains. These examinee proficiency measures are ordered with the highest values appearing at the top and the lowest at the bottom of the column. Each diamond represents one examinee.

The fourth column lists the five domains of the analytic scoring rubric, and the one holistic rubric, utilized in the scoring sessions in terms of their relative difficulty. More difficult scale categories appear higher in each column.

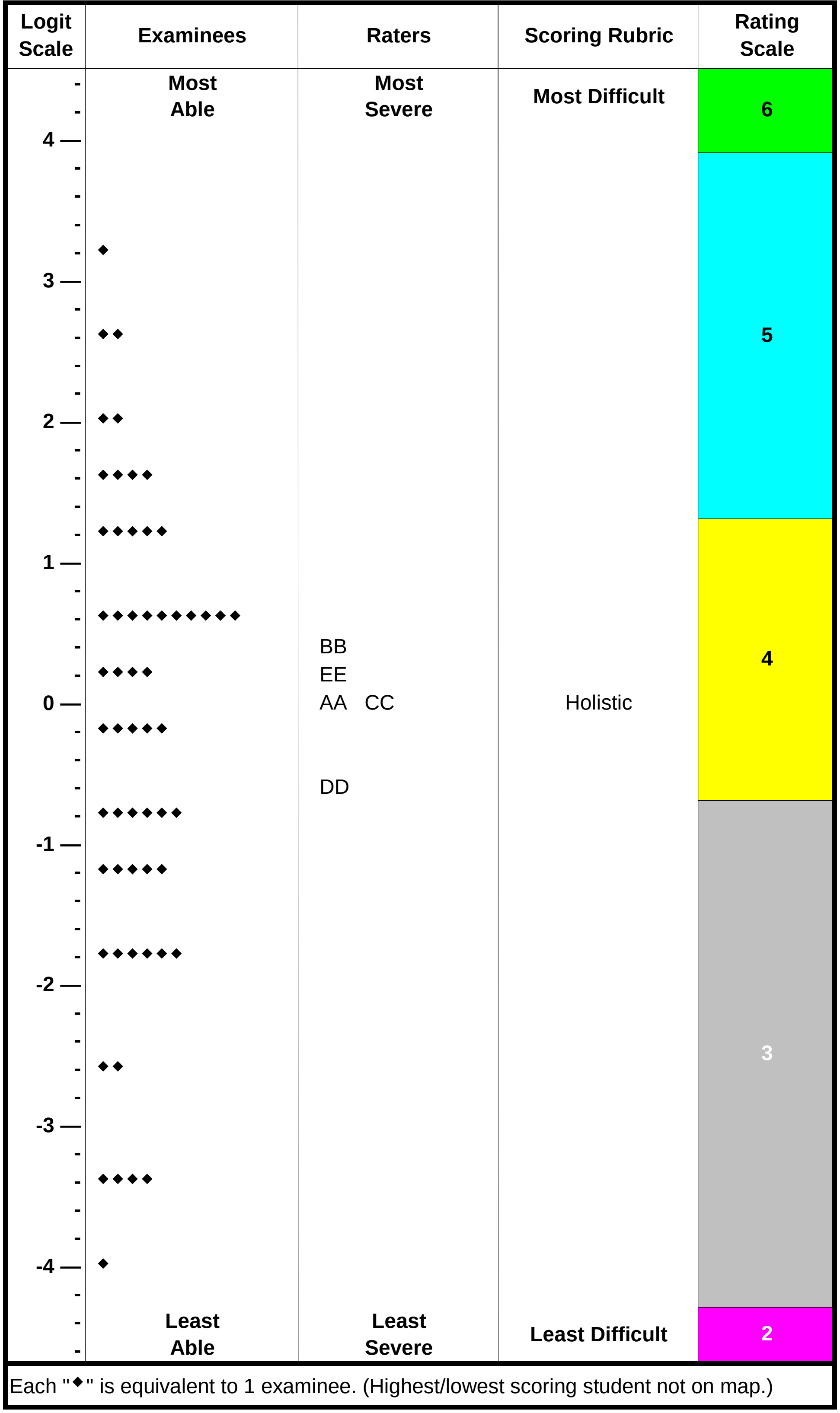
The last columns (five for the analytic scoring rubric and one for the holistic scoring rubric) display the six-point rating scale as used by raters to score the examinees on the analytic and holistic rubrics. The horizontal lines across these columns represent the point at which the probability of receiving the next higher rating begins to exceed the probability of receiving the lower rating. In the case of the task fulfillment analytic domain, for example, examinees with proficiency measures below –3.23 logits are most likely to receive a rating of 1; those with proficiency measures between –3.23 logits and –1.51 logits, a rating of 2; those with proficiency measures between –1.51 logits and 0.15 logits a rating of 3; those with proficiency measures between 0.15 logits and 1.73 logits a rating of 4; those with proficiency measures between 1.73 logits and 2.86 logits a rating of 5; and those with proficiency measures above 2.86 logits, a rating of 6.

Figure 1: Analytic Scoring Rubric Map



Each "◆" is equivalent to 1 examinee.

Figure 2: Holistic Scoring Rubric Map



Each "◆" is equivalent to 1 examinee. (Highest/lowest scoring student not on map.)

Table 2 provides a summary of statistics for ability estimates for the sixty examinees using both the analytic and holistic rubrics. When writing samples were scored with the holistic rubric, the mean examinee ability estimate was 3.8 with a standard deviation of 0.7; with the analytic rubric the mean was 3.5 with a standard deviation of 0.7. The same set of writing samples thus received a slightly higher score on average when scored using the holistic rubric as with the analytic rubric. The separation index and test reliability of examinee separation (the proportion of the observed variance in measurements of ability which is not due to measurement error) for the holistic scoring session were 2.31 and .84, respectively. For the analytic rubric the separation index and test reliability of examinee separation were 4.48 and. 95. This reliability statistic indicates the degree to which the analysis reliably distinguishes between different levels of ability among examinees. This measure is analogous to the KR20 index (Pollitt & Hutchinson, 1987). For examinees the reliability coefficient of .84 for examinees using the holistic rubric and .95 for the analytic rubric indicates that the analysis is fairly reliably separating examinees into different levels of ability. The chi-square of 316.40 for the holistically scored samples and 1186.00 for the analytically scored samples were both significant.

Table 2. Summary of Statistics on Examinees (N=60)

	Holistic Rubric	Analytic Rubric
Mean Ability	3.80	3.50
Standard deviation	.07	.07
Mean Square measurement error	.27	06
Separation Index	2.31	4.48
Test Reliability of examinee separation	.84	.95
Fixed (all same) chi-square	316.40	1186.00
	df=59, p <.001	df=59, p<.001

In order to identify examinees who exhibited unusual profiles of ratings for both the holistic and analytic rubrics, fit statistics presented in Tables 3 and 4 were examined. The tables show that four examinees (6, 36, 32, 9) showed misfit when the exam was analytically scored and six examinees (17, 42, 40, 51, 52, 6) showed misfit when the exam was scored holistically. There were, therefore, two more misfitting examinees when the writing samples were scored holistically. These findings would suggest taking a further look at the rating patterns for each of these examinees who showed misfit.

1. How much variability was there across examinee levels of proficiency?

The FACETS computer program calculated proficiency measures for each examinee in the sample. These measures were based on the probability of an examinee’s succeeding at a given level when scored on a particular domain, given that the domain was of a certain difficulty and that the raters involved in scoring the examinee essay were of varying levels of severity. As such, measures reflected the effects of the various conditions of assessment (examinee ability, rater severity, domain difficulty) on the score levels achieved (McNamara, 1996). In other words, examinee proficiency measures were calculated based on the assigned raw scores, but were adjusted for differences in rater severity and domain difficulty.

For the analytic domains, Table 3 and Figure 1 show that the distribution of examinee proficiency measures ranged from a low of -2.66 logits (SE = .31) to a high of 2.88 logits (SE = .27); a 5.54-logit spread. For the holistic scoring rubric, Table 4 and Figure 2 show that the distribution of examinee proficiency measures ranged from a low of -5.77 logits (SE = 1.11) to a high of 4.45 logits (SE = .86), a 10.22-logit spread. This meant that examinees were being separated across a wide range of proficiency levels. The range of examinee proficiency measures relative to the range of the rater severity was also important in determining the impact of individual differences in rater severity, as discussed below.

Table 3. Analytic scoring rubric- Examinee Summary Table (N=60)

Examinee	Observed Average	Fair Average	Measure (in logits)	S.E.	Infit Mean Square	Outfit Mean Square
28	5.1	5.10	2.88	0.27	0.70	0.67
21	4.8	4.86	2.45	0.26	1.14	1.05
53	4.8	4.78	2.32	0.26	0.70	0.68
22	4.7	4.70	2.19	0.25	0.32	0.34
29	4.7	4.70	2.19	0.25	1.04	1.05
49	4.6	4.61	2.07	0.25	0.63	0.60
38	4.6	4.57	2.00	0.25	1.03	1.09
14	4.4	4.45	1.82	0.25	0.47	0.49
10	4.4	4.37	1.69	0.25	1.35	1.36
57	4.4	4.37	1.69	0.25	1.47	1.45
1	4.3	4.28	1.57	0.25	0.54	0.54
48	4.1	4.08	1.27	0.25	1.12	1.16
55	4.1	4.08	1.27	0.25	0.80	0.80
56	4.0	4.00	1.14	0.25	1.30	1.31
13	3.9	3.91	1.02	0.25	1.00	0.99
50	3.9	3.91	1.02	0.25	0.66	0.67
2	3.9	3.87	0.95	0.25	0.82	0.83
31	3.8	3.83	0.89	0.25	0.58	0.58
47	3.8	3.83	0.89	0.25	0.75	0.73
24	3.8	3.75	0.76	0.26	0.76	0.74
44	3.7	3.71	0.69	0.26	0.75	0.73
51	3.7	3.71	0.69	0.26	1.08	1.10
6	3.7	3.67	0.63	0.26	2.44	2.29
17	3.6	3.59	0.49	0.26	0.99	0.91
25	3.6	3.59	0.49	0.26	1.10	1.07
39	3.6	3.59	0.49	0.26	0.58	0.60
52	3.6	3.59	0.49	0.26	1.61	1.54
30	3.5	3.51	0.35	0.26	1.38	1.36
41	3.5	3.51	0.35	0.26	0.53	0.54
16	3.5	3.47	0.28	0.27	0.88	0.90
7	3.4	3.43	0.21	0.27	0.70	0.65
37	3.4	3.43	0.21	0.27	0.89	0.87
5	3.4	3.39	0.14	0.27	0.89	0.85
46	3.4	3.35	0.07	0.27	0.92	0.83
54	3.4	3.35	0.07	0.27	0.66	0.65
19	3.3	3.31	0.00	0.27	1.15	1.12
59	3.3	3.31	0.00	0.27	0.47	0.47
40	3.3	3.27	-0.08	0.27	0.50	0.50
43	3.3	3.27	-0.08	0.27	0.22	0.23
32	3.2	3.23	-0.15	0.28	2.51	2.39
58	3.2	3.23	-0.15	0.28	1.08	1.08
18	3.2	3.19	-0.23	0.28	1.14	1.13
15	3.2	3.15	-0.30	0.28	0.65	0.59
45	3.1	3.11	-0.38	0.28	0.92	0.92
8	3.1	3.07	-0.46	0.28	1.05	1.10
4	3.0	3.03	-0.54	0.28	0.86	0.87
20	3.0	3.03	-0.54	0.28	0.69	0.70
60	3.0	3.00	-0.62	0.28	1.27	1.25
12	3.0	2.96	-0.70	0.29	1.13	1.17
27	2.9	2.92	-0.78	0.29	0.95	0.94
34	2.8	2.76	-1.11	0.29	0.67	0.69
33	2.7	2.72	-1.20	0.29	1.24	1.20
35	2.6	2.64	-1.37	0.29	0.86	0.89
9	2.6	2.60	-1.46	0.30	2.89	2.88
3	2.6	2.56	-1.55	0.30	1.62	1.57
42	2.6	2.56	-1.55	0.30	0.91	0.92
26	2.4	2.44	-1.81	0.30	0.98	0.94
11	2.3	2.32	-2.09	0.30	0.66	0.66
23	2.2	2.13	-2.53	0.35	0.50	0.52
36	2.1	2.08	-2.66	0.31	2.23	2.20
Mean	**3.5**	**3.51**	**0.26**	**0.27**	**1.00**	**0.98**
SD	**0.7**	**0.70**	**1.24**	**0.02**	**0.51**	**0.49**

Table 4. Holistic scoring rubric- Examinee Summary Table (N=60)

Examinee	Observed Average	Fair Average	Measure (in logits)	S.E.	Infit Mean Square	Outfit Mean Square
28	5.6	5.60	4.45	0.86	0.94	0.91
21	5.2	5.19	3.17	0.76	0.50	0.51
48	5.0	4.99	2.61	0.74	1.73	1.73
49	5.0	4.99	2.61	0.74	0.05	0.05
22	4.8	4.79	2.08	0.72	0.29	0.28
43	4.8	4.79	2.08	0.72	0.29	0.28
10	4.6	4.60	1.58	0.70	0.82	0.83
38	4.6	4.60	1.58	0.70	0.46	0.45
47	4.6	4.60	1.58	0.70	1.34	1.33
52	4.6	4.60	1.58	0.70	2.33	2.26
1	4.4	4.40	1.10	0.68	0.53	0.53
41	4.4	4.40	1.10	0.68	0.44	0.44
53	4.4	4.40	1.10	0.68	0.44	0.44
56	4.4	4.40	1.10	0.68	0.53	0.53
57	4.4	4.40	1.10	0.68	0.35	0.35
5	4.2	4.20	0.64	0.68	0.44	0.44
6	4.2	4.20	0.64	0.68	2.07	2.06
24	4.2	4.20	0.64	0.68	0.42	0.42
25	4.2	4.20	0.64	0.68	0.16	0.17
29	4.2	4.20	0.64	0.68	0.16	0.17
39	4.2	4.20	0.64	0.68	1.25	1.25
44	4.2	4.20	0.64	0.68	0.16	0.17
46	4.2	4.20	0.64	0.68	0.42	0.42
50	4.2	4.20	0.64	0.68	0.99	0.99
51	4.2	4.20	0.64	0.68	2.92	2.90
14	4.0	4.00	0.19	0.68	0.89	0.89
37	4.0	4.00	0.19	0.68	0.89	0.90
45	4.0	4.00	0.19	0.68	0.63	0.63
55	4.0	4.00	0.19	0.68	1.73	1.74
13	3.8	3.80	-0.27	0.69	0.27	0.28
17	3.8	3.80	-0.27	0.69	4.38	4.37
32	3.8	3.80	-0.27	0.69	0.27	0.28
58	3.8	3.80	-0.27	0.69	1.02	1.01
59	3.8	3.80	-0.27	0.69	0.35	0.36
2	3.6	3.60	-0.76	0.71	1.15	1.11
9	3.6	3.60	-0.76	0.71	0.77	0.75
19	3.6	3.60	-0.76	0.71	1.41	1.41
30	3.6	3.60	-0.76	0.71	1.45	1.45
34	3.6	3.60	-0.76	0.71	1.59	1.63
54	3.6	3.60	-0.76	0.71	0.65	0.66
4	3.4	3.40	-1.29	0.75	0.99	1.01
12	3.4	3.40	-1.29	0.75	0.54	0.55
20	3.4	3.40	-1.29	0.75	0.47	0.47
42	3.4	3.40	-1.29	0.75	3.49	3.37
60	3.4	3.40	-1.29	0.75	0.46	0.45
7	3.2	3.20	-1.90	0.81	0.74	0.78
8	3.2	3.20	-1.90	0.81	0.29	0.26
15	3.2	3.20	-1.90	0.81	0.74	0.78
16	3.2	3.20	-1.90	0.81	1.99	2.09
27	3.2	3.20	-1.90	0.81	0.29	0.26
31	3.2	3.20	-1.90	0.81	0.74	0.78
40	3.2	3.20	-1.90	0.81	3.42	3.34
18	3.0	3.01	-2.59	0.86	1.15	1.13
35	3.0	3.01	-2.59	0.86	0.04	0.04
3	2.8	2.81	-3.32	0.85	0.62	0.63
23	2.8	2.81	-3.32	0.85	0.88	0.90
26	2.8	2.81	-3.32	0.85	0.54	0.53
33	2.8	2.81	-3.32	0.85	0.62	0.63
11	2.6	2.61	-4.04	0.84	0.80	0.79
36	2.2	2.20	-5.77	1.11	1.18	1.48
Mean	**3.8**	**3.85**	**-0.30**	**0.74**	**0.96**	**0.96**
SD	**0.7**	**0.69**	**1.88**	**0.08**	**0.87**	**0.86**

2. To what extent did the scoring rubrics succeed in separating examinees into distinct strata of proficiency? How many statistically different levels of examinee proficiency were there?

The FACETS computer program calculated an examinee separation ratio (G), which was a measure of the true spread of examinee proficiency estimates relative to their measurement error (Linacre, 2005). This separation ratio was used in the formula $\frac{4G+1}{3}$ to calculate the number of statistically distinct proficiency strata into which the test separated the examinees (Myford & Wolfe, 2000).

The 60 examinees were separated into approximately six statistically distinct levels of proficiency (G = 4.48, Strata ≅ 6) when the writing samples were scored via the analytic scoring rubrics. This variability was statistically significant, χ^2 (59, N = 60) = 1186.0, p < .001. The holistic scoring rubric separated the 60 examinees into approximately 3 statistically distinct levels of proficiency (G = 2.31, Strata ≅ 3). This variability was statistically significant, χ^2 (59, N = 60) = 316.4, p < .001. That is, after allowing for measurement error, the probability that all the examinees in the sample had the same level of proficiency was practically nil.

3. Were the differences between examinee proficiency due mostly to measurement error or to actual differences in proficiency?

The reliability with which a test separates a sample of examinees is calculated as the ratio between the variance due to the construct being measured (true score variance) and the observed variance (true score variance plus error variance) (Wright & Masters, 1982). The person separation reliability produced by the FACETS program is a measure of how different the examinee proficiency measures are in contrast to the inter-rater reliability coefficient produced by FACETS, which is a measure of how similar rater measures are (Linacre, 2005).

For the analytic scoring rubric, the examinee separation reliability coefficient was .95, and for the holistic scoring rubric the examinee separation reliability coefficient was .84. These high reliability values suggested that the true variance exceeded the error variance in the examinee proficiency measures. The two scoring sessions thus succeeded in distinguishing between the different levels of examinee proficiency, although not equally.

5.3 Rubric Domains

1. To what extent did the domains in the analytic rubric differ in their levels of difficulty? Was it harder for examinees to get high ratings on some domains than on others?

Table 5 lists the five analytic domains used in the analytic scoring session in difficulty order. The domains ranged in difficulty from -0.10 logits (SE = .08) to 0.12 logits (SE = .07); a .22-logit span. The average scoring domain difficulty was .00 with a corresponding measurement error of .08. The reliability of separation for the domains in the analytic scoring rubrics was .00, which suggested that these scoring domains functioned interchangeably. The separation index of .00

provided further evidence of interchangeability, indicating that there were no differing levels of difficulty. The overall difference between the difficulty of these five domains was not statistically significant χ^2 (4, N = 5) = 4.9, p > .05. That is, there was a high probability that these domains were in fact similar in difficulty.

Table 5. Analytic scoring rubric summary

Domain	Observed Average	Fair Average	Measure (in logits)	S.E.	Infit Mean Square	Outfit Mean Square
Topic Development	3.4	3.29	0.12	0.07	1.06	1.06
Task Fulfillment	3.6	3.56	0.04	0.07	1.04	1.04
Linguistic Control	3.5	3.49	0.00	0.08	1.00	1.00
Organization	3.5	3.45	-0.06	0.08	0.94	0.94
Vocabulary & Register	3.5	3.46	-0.10	0.08	0.84	0.87
Mean	**3.5**	**3.45**	**0.00**	**0.08**	**0.98**	**0.98**
SD	**0.1**	**.09**	**0.08**	**0.00**	**0.08**	**0.07**

The fit indices for all five analytic scoring domains were within the desired bounds of .7 to 1.4. Furthermore, both infit mean-square and outfit mean-square values were very close to the expected value of 1.0 indicating that the rating patterns for each of the scoring domains were very close to those expected by the FACETS model.

Therefore, I concluded that it was not harder for examinees to get high ratings on some analytic domains than on others. The domains of the analytic scoring rubric were nearly equal in difficulty, and as such could be said to function interchangeably. Also, the scoring pattern for each domain fit the model expectation.

5.4 Rating Scale

1. Were the six categories of the analytic and holistic rating scales appropriately ordered? Were the analytic and holistic rating scales functioning properly as six-point scales? Were the scale categories clearly distinguishable, i.e., "most probable" over clearly defined intervals?

To answer these questions, I examined the average examinee proficiency measures, the outfit mean-square index, and the category probability curves, for each of the domains under investigation. The average examinee proficiency measure for a particular category was obtained by averaging the proficiency measures of all the examinees that received a rating in that category for a given essay. The rating scale functions as expected when the average examinee proficiency measures increases in value as the rating scale categories advance (Linacre, 2005).

Table 6 lists the average measures for the six categories used in scoring the five analytic domains. I observed that, for nearly all domains, the average examinee proficiency measures increased in magnitude as the rating scale categories increased. This suggested that, on average, examinees with higher ratings on a particular domain were indeed more able than examinees with lower ratings on the same domain.

Table 6. Analytic scoring rubric – Average examinee ability measures

Category	Domain				
	Task Fulfillment	Topic Development	Organization	Vocabulary & Register	Linguistic Control
1	-0.81	-1.92	-1.47	-1.67	-1.56
2	-1.10	-0.96	-1.03	-1.22	-1.12
3	-0.40	-0.16	-0.20	-0.14	-0.24
4	0.53	0.55	0.71	0.84	0.65
5	1.50	1.47	1.73	1.90	1.65
6	2.38	2.41	2.66	3.06	2.41

The only exception occurred in the case of the Task Fulfillment domain for which the average examinee proficiency measure for category 2 (-1.10 logits) was lower than the average examinee proficiency measure for category 1 (-.81 logits). This was an unexpected result, which suggested that the proficiency measures of at least some of the examinees receiving “1”s on this domain may have been less accurately determined. It is important to note, however, that category 1 of the Task Fulfillment domain was used only 3% of the time, and as such it may be imprecisely estimated and unstable (Linacre, 1999). When fewer than 10 ratings are present in any category, the calculation of the average examinee proficiency measures for that category is less accurate. Since raters infrequently used the 1st category of the Task Fulfillment domain, I might attribute this skip in the expected advance of examinee average measure to a statistical aberration, and conclude that, overall, the average measures increased as expected.

Table 7 lists the average measures used for the holistic scoring rubric. The average examinee ability measure increased as expected for categories 2 through 6. However, no such value was present for category 1. Raters did not employ this category at all. It appeared that when assigning scores using the holistic scoring rubric, these raters did not come across any essay deserving of a score of 1. The holistic rubric, therefore, appeared to be functioning not as a six-point but as a five-point scale.

Table 7. Holistic scoring rubric – Average examinee ability measures

Category	Holistic Scoring Rubric
1	-
2	-3.33
3	-1.7
4	.10
5	1.46
6	2.55

Another useful indicator of rating scale functionality is the outfit mean-square statistic. This statistic has an expected value of 1 and was calculated for each rating scale category. When the observed and expected examinee proficiency measures are close, the outfit statistic is close to 1.0. When the discrepancy between observed and expected average measures is large, the outfit statistic will be large as well. Outfit values less than .5 suggest over-predictability and may bias the computations of reliability and separation, while values greater than 2.0 indicate that there is more "noise" (unexplained variability) in the ratings than statistical information (Linacre, 1999).

All but one outfit mean-square indices for the analytic domains presented in Table 8 were close to the expected value of 1.0. The only exception was again encountered in the case of category 1 of the Task Fulfillment domain. The fact that this category was only used eight times clearly impacted not only the calculations of the examinee average measures, but also the fit statistics associated with it. The high value of this outfit statistic might be explained by instability caused by the low number of ratings present in category 1 of this domain.

Table 8. Analytic scoring rubric – Outfit mean-square indices

Category	Task Fulfillment	Topic Development	Organization	Vocabulary & Register	Linguistic Control
1	2.2	1.1	1.3	1.1	1.3
2	1.0	1.1	1.1	0.9	1.1
3	1.1	1.3	0.9	0.9	0.9
4	1.0	1.1	1.0	0.9	1.0
5	0.8	0.9	0.8	0.9	1.0
6	0.8	0.7	0.7	0.7	0.8

As seen in Table 9, the outfit mean-square indices for the holistic scoring rubric were also close to the expected value of 1.0. This suggested that the 5 categories that were used during the holistic scoring session functioned as expected.

Table 9. Holistic scoring rubric – Outfit mean-square indices

Category	Holistic Scoring Rubric
1	-
2	1.2
3	0.9
4	0.8
5	0.9
6	1.4

Rating scale category thresholds provided another indicator of rating scale functionality. The Partial Credit Model (Equation 1) used to analyze these data allows the rating scale category structure to vary from one domain to another. This allowed us to examine the rating scale thresholds for each domain. A rating category threshold is the point on the examinee proficiency scale at which probability curves for adjacent categories intersect. At this point, an examinee has a 50% chance of being rated in either of the two adjacent categories, given that the examinee is rated on one of them (Andrich, 1998). Increasing category thresholds indicate a well functioning rating scale, in which each category is adequately used, (i.e., it has a chance of being "most probable.")

Table 10 presents the rating scale thresholds for the analytic domains. For each of the five domains, the thresholds advanced monotonically as desired, that is, the number in each column increased without any two numbers being equal. The distance between thresholds in each domains was also adequate; that is, for each category in each domain, there was enough of a difference in the value of each threshold so that each category has a chance of being "most probable." A graphical interpretation of this statement can be seen in Figures 3 – 7 below. When a particular rating scale category functions adequately, its graph will "peak" between intersection points (thresholds) on the probability curves graph. This is clearly the case for each analytic domain.

Table 10. Analytic scoring rubric – Rating scale thresholds

Category Threshold	Task Fulfillment	Topic Development	Organization	Vocabulary & Register	Linguistic Control
1 - 2	-3.23	-4.02	-4.12	-5.20	-3.90
2 - 3	-1.51	-1.36	-1.81	-1.83	-1.75
3 - 4	0.15	0.80	0.67	0.63	0.37
4 - 5	1.73	1.59	1.89	2.12	2.03
5 - 6	2.86	2.98	3.37	4.28	3.24

Figure 3. Category Probability Curves – Task Fulfillment

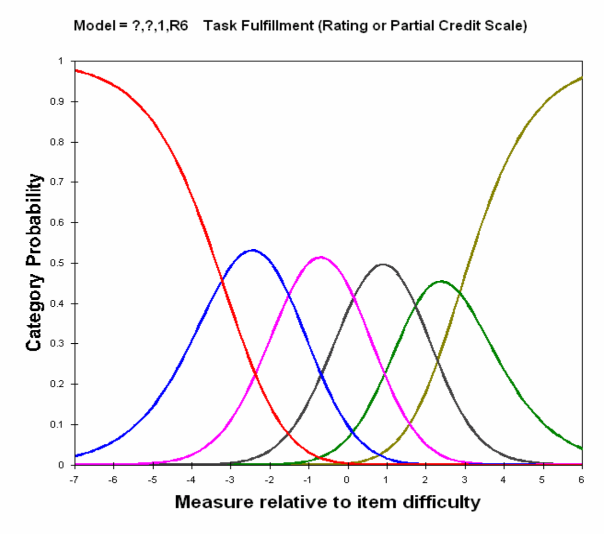


Figure 4. Category Probability Curves – Topic Development

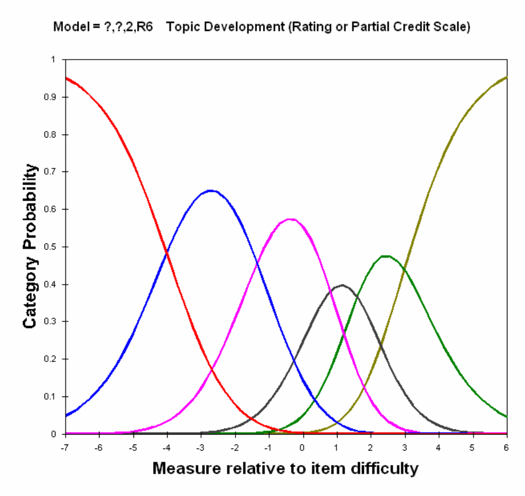


Figure 5. Category Probability Curves – Organization

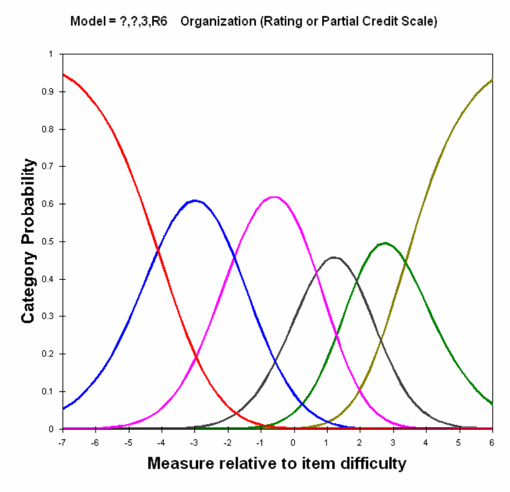


Figure 6. Category Probability Curves – Vocabulary & Register

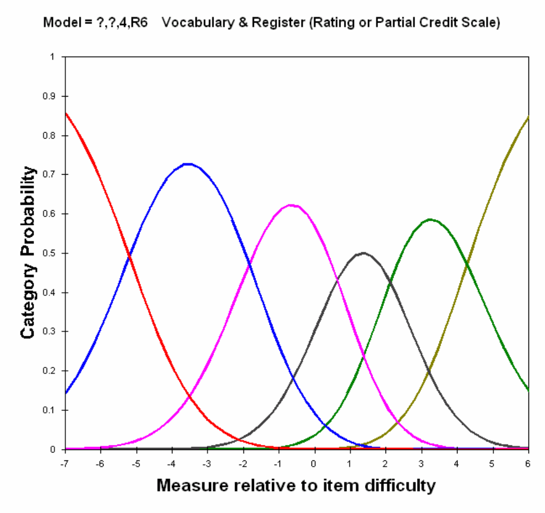
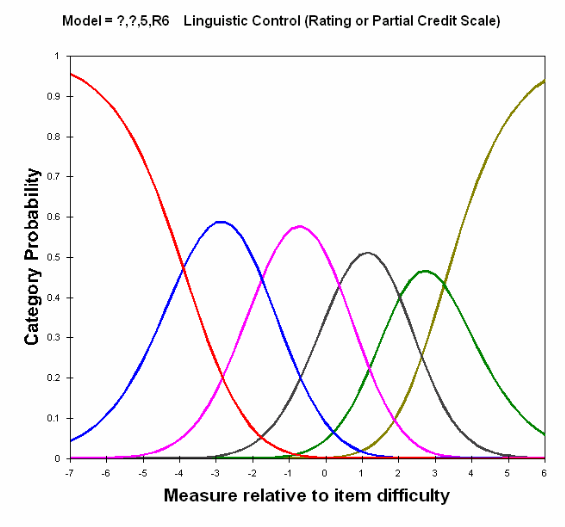


Figure 7. Category Probability Curves – Linguistic Control



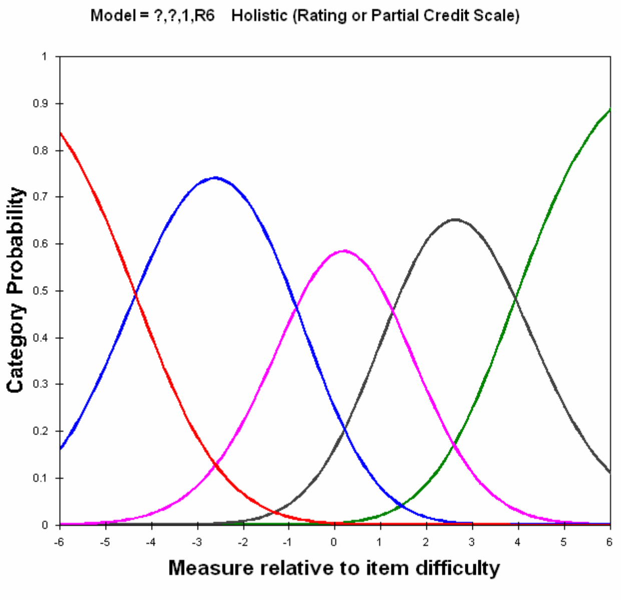
A close inspection of the thresholds listed in Table 11 shows that they varied across domains. When taking into account the SE of the threshold values, we note that the threshold between categories 3 and 4 varied significantly between Task Fulfillment and Topic Development, Organization, and Vocabulary and Register. This particular threshold was also found to vary significantly between Topic Development and Linguistic Control. The only other significant difference could be found for the threshold between categories 5 and 6, which occurred between Vocabulary and Register, and Task Fulfillment and Topic Development. These significant differences suggested that the categories adjacent to these thresholds were used differently in the domains in question (i.e., the involved categories were differentially "most probable" for the various domains in question).

Category thresholds for the holistic scoring rubric also increased as expected. As seen in Figure 8, the category probability curves peaked over sufficiently large intervals, indicating that, with the exception of the missing category 1, this scale functioned adequately.

Table 10. Holistic scoring rubric – Rating scale thresholds

Category Threshold	Holistic Scoring Rubric
1 - 2	-3.23
2 - 3	-1.51
3 - 4	0.15
4 - 5	1.73

Figure 8. Category Probability Curves – Holistic Scoring Rubric



In sum, the monotonically advancing average measures and category thresholds indicated that the rating scales functioned properly for every essay sample. However, while the rating scale categories were clearly distinguishable, category functioned differentially for some of the domains of the analytic scoring rubric.

6. Discussion

The analyses revealed a well spread-out distribution of student proficiency measures for both the analytic and holistic scoring rubrics. Both scoring rubrics were able to effectively spread out the examinees into distinct strata of ability. The analytic scoring rubrics produced 6 statistically distinct levels of proficiency. This was equal to the number of categories used to score the analytic scoring rubric domains. The holistic scoring rubric was less effective at separating examinees into distinct groups, as it only produced 3 strata. Since only 5 of the 6 available categories were actually employed in the holistic scoring session, it is expected that it would be less discriminating. The analytic scoring rubric was thus more sensitive to distinguishing the differing levels of ability present in the sample of examinees.

The five domains of the analytic scoring rubric – Task Fulfillment, Topic Development, Organization, Vocabulary & Register, and Linguistic Control – did not differ in difficulty. The difficulty distribution covered a very small range (.22 logits) indicating that these domains functioned interchangeably. In addition, the scoring patterns obtained for each domain fit model expectations very closely. As a result, it was not harder for examinees to obtain high ratings on some analytic domains than on others.

The rating scale of the analytic scoring rubric functioned as expected in each domain. The advancing average examinee measures indicated that examinee proficiency increased as the category value increased. In other words, higher scores were assigned to examinees of higher proficiency.

The infit mean square index for the analytic domains also performed as expected with all fit statistics well within the pre-established bounds. The adequate infit statistics suggested the presence of very little discrepancy between the observed and expected average measures, (i.e., neither too much over-predictability, nor too much unexplained variability was detected in the data). The third indicator of rating scale functionality, advancing thresholds, was also found to be present in the data. Each category was "most probable" over a distinct, non-zero interval, indicating that the raters made use of each category when rating the examinees' essays. Only two minor discrepancies were found. Categories 1 and 2 of the Task Fulfillment domain were employed for only a small percentage of times and as a result, the average measure for category 2 and outfit statistic for category 1 were slightly outside expected norms. Otherwise, the categories of the analytic scoring rubric functioned properly as a six-point scale.

The functioning of six categories, however, was not the same across all domains. Significant differences between the thresholds of category 3 and 4, and 5 and 6 were found when comparing the rating scale performance across domains. This is an indicator that some categories were more likely to be used for certain domains than for others. Category 4, for example is most probable for only 0.79 logits for the Topic Development domain, but is most probable for 1.66 logits for the Linguistic Control domain. Another significant difference can be noted in the case of category 5 for Vocabulary & Register (most probable over 2.16 logits) and Linguistic Control (most probable over 1.21 logits). It appears that while all six categories were used for every domain, some categories were more preponderant than others. It would be of interest to further investigate if the meaning of some of these categories changes between domains, if raters are more inclined to use a particular category in a particular domain, or if these differences indeed reflected differences in the various dimensions of L2 writing ability, feedback that could inform instruction in the ESL writing curriculum.

The rating scale of the holistic scoring rubric functioned adequately, but not as expected. Although average measures increased as category numbers increased, outfit values were within bounds, and thresholds advanced strictly monotonically, the scale only functioned as a five-point scale. Category 1 was never present in the ratings obtained in this session. This suggests that raters may have been reluctant to assign the lowest rating of 1. It could have been that they avoided giving ratings of 1 in general or perhaps they avoided giving a rating of 1 to writing samples submitted on final exit examinations. The sample of examinees under analysis in this scoring session was limited to 60. It could be that a significantly larger sample would produce essays that only deserve a score of 1, but it is not likely that this category will be over-used even in a larger sample. The performance of the holistic rubric as a 5-point scale suggested that the holistic rubric might perform better if category 1 and 2 were collapsed to create a 5-point scale or it might be that rater training and norming sessions include more samples that deserve a 1 to encourage the use of that category as a valid score.

In sum, it can be said that of the two rubrics, the analytic scoring rubric functioned better as a six-point scale than the holistic scoring rubric. Overall, the analytic rubric was more discriminating than the holistic rubric although discrepancies between the differential uses of the six categories among domains warrant further investigation.

While this study provided evidence of reliability and validity for both the holistic and analytic rubrics, it suggested that the analytic rubric might better separate examinees over a greater number of strata. That said, the additional information provided by the analytic rubrics (5 scores as opposed to 1) could produce a profile of the L2 writing ability of a test-taker, which would be valuable for diagnostic and placement purposes, given that different aspects of L2 writing ability develop at different rates for different writers.

Even though scoring holistically often translates into a savings in time and money in the short term, an analytic rubric, such as the one developed for this assessment, may represent a long-term savings, both for the student and the program and thus be preferable for diagnostic and/or placement purposes. The individual scores assigned for different domains of writing ability using the analytic scoring method provide additional information to reduce the incidence of misclassifications. Accurate placement results in more homogeneous groups of learners and a more effective learning environment. With a more complete profile of each student's writing ability, the teacher can adjust curricular objectives accordingly. The performance of these two rubrics in the assessment of L2 writing suggested that the analytic rubric with five domains should be adopted to provide critical information for accurate placement decisions and valuable diagnostic information about individual student's L2 writing ability to adjust curricular goals and objectives.

References

Andrich, D. (1988). *Rasch models for measurement.* Newbury Park: Sage.

Andrich, D. (1998). Thresholds, steps, and rating scale conceptualization. *Rasch Measurement:* Transactions, 12(3), 648.

Alloway, J.E. (1978). Some ways of establishing criteria for assessing writing performance from the perspective of the test developer. Paper presented at the annual meeting of the American Education Research Association, Toronto. Cited in Smith, et al (1980).

Bachman, L. & Palmer, A. (1996). *Language testing in practice: designing and developing useful language tests*. Oxford: Oxford University Press.

Bachman, L. & Savignon, S.J. (1986). The evaluation of communicative language proficiency: a critique of the ACTFL Oral Interview. *Modern Language Journal*, 70(4), 380-90.

Barkaoui, K. (2010). Explaining ESL essay holistic scores: A multilevel modeling approach. *Language Testing*, 27(4), 515-535.

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

Cohen, L. & Manion L. (1994). *Research methods in education*. New York: Routledge

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

Elbow, P. (1999). Ranking, evaluating, and liking: sorting out three forms of judgments. In R. Straub (Ed.), *A sourcebook for responding to student writing* (pp. 175-166). New Jersey: Hampton Press.

Engelhard, G., Jr., and Myford, C. (2003). *Monitoring faculty consultant performance in the advanced placement English literature and composition program with a many-faceted Rasch mode*l. (College Board Research Report No 2003-1). New York, NY: College Entrance Examination Board.

Fulcher, G. (1987). Tests of oral performance: The need for data-based criteria. *English Language Teaching Journal*, 14(4), 287-291.

Godshalk, F.I., Swineford, F. & Coffman, W.E. (1966). The measurement of writing ability. New York: College Entrance Examination Board.

Hamp-Lyons, L. (2003). Writing teachers as assessors of writing. In Kroll, B. (Ed.), Exploring the dynamics of second language writing (pp.162-189). New York: Cambridge University Press.

Hamp-Lyons, L. (1991). Scoring procedures for ESL context. In L. Hamp-Lyons (Ed.), Assessing second language writing (pp. 241-277). Norwood, NJ: Ablex.

Hamp-Lyons, L. (1990). Second language writing: Assessment issues. In B. Kroll (Ed.), Second language writing: Research insights for the classroom (pp. 69-87). New York: Cambridge University Press.

Huang, J. (2008). How accurate are ESL students' holistic scores on large-scale assessments? -- A generalizability theory approach. Assessing Writing, 13, 201-218.

Hyland, K. (2002). Teaching and research Writing. Harlow, England: Pearson Education Limited.

Jacobs, H., Zingraf, A., Warmuth, D., Hartfiel, V. & Hughey, J. (1981). Testing ESL composition: A practical approach. Rowley, MA: Newbury House.

Janopoulos, M. (1993). Comprehension, Communicative Competence, and Construct Validity: Holistic scoring from an ESL perspective. In M.M. Williamson & B.A. Huot (Eds.), Validating holistic scoring for writing assessment: Theoretical and empirical foundations (pp. 303-325). Cresskill, NJ: Hampton Press, Inc.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. Language Testing, 26(2), 275-304.

Kroll, B. (1990). Second language writing: Research insights for the classroom. New York: Cambridge University Press.

Linacre, J.M. (2005). A user's guide to FACETS: Rasch measurement computer program. Version 3.57. Chicago, IL.

Linacre, J.M. (1999). Investigating rating scale category utility. Journal of Outcome Measurement, 3(2), 103-122.

Linacre, J.M. (1989). Many-facet Rasch measurement. Chicago: MESA Press.

Matthews, M. (1990). The measurement of productive skills: Doubts concerning the assessment criteria of certain public examinations. ELT Journal, 44(2), 117-21.

McNamara, T. (1996). Measuring second language performance. New York: Addison Wesley Longman.

Myford, C. & Wolfe, E.W. (2000). Monitoring sources of variability within the Test of Spoken English assessment system (Research Project 65). Princeton, NJ: Educational Testing Service.

Perkins, K. (1983). On the use of composition scoring techniques, objective measure, and objective tests to evaluate ESL writing ability. TESOL Quarterly, 17(4), 651-71.

Pollitt, A. & Hutchinson, C. (1987). Calibrated graded assessments: Rasch partial credit analysis of performance in writing. Language Testing 4(1), 72-92.

Powills, J.A., Bowers, R. & Conlan, G. (1979). Holistic essay scoring: An application of the model for the evaluation of writing ability and the measurement of growth in writing ability over time. Paper presented at the annual meeting of the American Education Research Association, San Francisco.

Rasch, G. (1980). Probabilistic models for some intelligence and attainment tests. Chicago: The University of Chicago Press.

Reid, J. (1993). Teaching ESL Writing. Englewood Cliffs, NJ: Regents Prentice Hall.

Shaw, S.D. (2002). IELTS Writing: revising assessment criteria and scales (Phase 2). Research Notes, 10, 10-12.

Shohamy, E. (1995). Performance assessment in language testing. Annual Review of Applied Linguistics, 15, 188-211.

Smith, L.S., Winters-Edys, L., Quellmalz, S, & Baker, E.L. (1980). Characteristics of student writing competence: An investigation of alternative scoring systems. CSE Report No. 134. Center for the Study of Evaluation: UCLA. ERIC 217074

Upshur, J.A. & Turner, C.E. (1995). Constructing rating scales for second language tests. ELT Journal, 49(1), 3-12.

Weigle, S.C. (2002). Assessing Writing. Cambridge: Cambridge University Press.

White, E.M. (1985). Teaching and assessing writing. San Francisco, CA: Jossey-Bass.

Williamson, M.M. & Huot, B.A. (1993). Validating holistic scoring for writing assessment: Theoretical and empirical foundations. Cresskill, NJ: Hampton Press, Inc.

Wright, B.D. & Masters, G.N. (1982). Rating scale analysis, Rasch Measurement. Chicago, IL: Mesa Press.

Appendix A

Holistic Scoring Scale

Criteria for Grading ESL Papers

6

The essay shows a basic understanding of the demands of essay organization although there might be occasional digressions. The development of ideas is sometimes incomplete or rudimentary, but a basic logical structure can be discerned. Vocabulary generally is appropriate for the essay topic but at times is oversimplified. Sentences reflect a sufficient command of standard written English to ensure reasonable clarity of expression. Common forms of agreement and grammatical inflection are usually, though not always, correct. The writer generally demonstrates through punctuation an understanding of the boundaries of the sentences. The writer spells common words, except perhaps so-called “demons,” with a reasonable degree of accuracy.

5

The focus of the essay is clear but may have a few digressions. The examples used, particularly those from personal experience, are rich, and the writer may occasionally cite statistics or make reference to personal readings. However, examples may not be entirely relevant or appropriate for the topic. The essay is effectively organized, but the basic logical structure may be flawed. The writer demonstrates good control of basic and complex patterns of sentence construction but usually the variety is rather limited. Vocabulary is generally adequate for the topic, and the writer demonstrates basic control of word choices, word forms and idiomatic expressions. Although the essay reflects the above qualities, it is not error-free and has a non-native flavor.

4

In general the focus of the essay is clear despite some digressions and contradictions. The writer makes an attempt at sophistication in supporting details. For the most part, there is a logical connection between the introduction, body and conclusion. The essay sounds finished. The writer demonstrates good control of basic patterns of sentence construction and uses some complex structures though not always correctly. The essay contains lingering errors or interferences that may seem out of place given length, development and ease of expression. The writer may attempt to use rhetorical devices such as deliberate repetition, contrast, parallelism and so forth. The writer uses linking devices although by no means in firm control. Conventions are more often correct than incorrect. Vocabulary is varied, but there are some inappropriate word choices and word forms.

3

The essay lacks focus in development of the central idea. The paper may be a page and a half or more in length. The writer makes an attempt at development although examples are often irrelevant. Paragraphs are more than two sentences long, and there are usually three or four of them; there is an introduction, a body and a conclusion. The essay contains occasional first language interference. The writer attempts to use transitional devices. The essay contains many run-ons rather than fragments. It is repetitious and contains many misspelled words. Communication is not lost despite many errors in grammar.

2

The paper may be one page or more in length. It contains many fragments. The writer makes some attempt at discussing ideas and shows some evidence of organization of ideas (paragraphs are often one sentence). There is much first language interference, often due to direct translation. The writer makes an attempt at expressing fairly complex ideas and using tenses although usually incorrect in usage and form. The errors in grammar occur so frequently that communication is lost, and there are many misspelled words.

1

The paper is very short, usually a half page to one page, and is often incomprehensible. There is strong evidence of first language interference. The paper is sometimes written completely in the writer's first language. Sentence structure is simple, and vocabulary is very limited. The spelling is poor, and there is little use of correct punctuation. Control of grammar is weak, and the paper lacks fluency.

Appendix B

Analytic Scoring Rubric

	1 point	2 points	3 points	4 points	5 points	6 points
Task Fulfillment	• Fails in attempt to write an agree/disagree essay, or • Fails in attempt to write a narrative.	• Attempts to write an agree/disagree essay, or • Attempts to write a narrative	• Minimally succeeds in writing an essay, or • Minimally succeeds in writing a descriptive narrative.	• Writes an essay that adequately expresses a point of view, or • Writes an essay that adequately develops a descriptive narrative.	• Writes an essay that competently expresses a point of view, or • Writes an essay that competently develops a descriptive narrative.	• Writes an essay that convincingly expresses a point of view, or • Writes an essay that fully develops a descriptive narrative.
Topic Development	▪ Fails to provide related support	▪ May provide 1-2 points directly or indirectly related to topic ▪ Limited development of topic	▪ Provides 1-2 points mostly related to topic with occasional digressions ▪ Provides some development of topic	▪ Provides 2+ points that directly relate to topic ▪ Adequate development of topic	▪ Provides 2+ points that adequately support topic ▪ Substantial development of topic	▪ Provides 2+ convincing points related to topic ▪ Thorough development of topic
Organization	▪ Follows no clear organizational pattern	▪ May lack clear organizational pattern ▪ Mechanical or inaccurate attempts at cohesive devices	▪ Discernable organizational pattern (listing ideas, etc) ▪ Rudimentary use of cohesive devices	▪ Generally organized ▪ Some use of cohesive devices	▪ Well organized ▪ Demonstrates competent use of cohesive devices	▪ Focused and organized ▪ Demonstrates skillful use of cohesive devices
Register & Vocabulary	▪ Demonstrates limited vocabulary ▪ Generally inaccurate &/or inappropriate word choice or register	▪ Demonstrates minimal range of vocabulary ▪ Often uses inappropriate word choice and/or register	▪ Demonstrates narrow range of vocabulary ▪ Some inaccurate or inappropriate word choices or register	▪ Demonstrates adequate range of vocabulary ▪ Occasional inaccurate or inappropriate word choices or inappropriate register	▪ Demonstrates competent range of vocabulary for social and academic purposes ▪ Generally accurate word choice but not always appropriate usage ▪ Demonstrates ability to write in academic register	▪ Demonstrates extensive range of vocabulary for social and academic purposes ▪ Few problems with word choice ▪ Demonstrates ability to write in appropriate academic register
Linguistic Control	▪ Frequent errors of all types with little control ▪ Errors generally obscure meaning ▪ Lacks basic	▪ Frequent errors with uneven control of language ▪ Errors often obscure meaning ▪ Lacks	▪ Frequent errors but demonstrates minimal control of language ▪ Errors sometimes interfere with	▪ May make frequent errors but demonstrates developing grammatical control ▪ Errors occasionally	▪ Some errors but control of language is apparent ▪ Few, if any, errors that interfere with meaning ▪ Demonstrates	▪ A few errors are noticeable ▪ Errors rarely interfere with meaning ▪ Variety of simple and complex

	sentence structure & variety	sentence variety – mostly simple complete sentence	meaning ▪ Limited sentence variety – simple and some compound or occasional complex sentence	interfere with meaning ▪ Demonstrates some sentence variety with simple, compound and some complex sentences but with errors, e.g., fragments, run-ons, errors in subordination	sentence variety but with some errors	sentence structures with few errors