国際研究論叢　24（2）：77～95，2011

A FACETS Analysis of the JACET Kansai Working Group Writing Rubric 2009

Brad Visgatis *

Abstract

The purpose of this paper is to use Rasch analyses to examine the 2009 version of a writing rubric developed by the JACET Kansai working group and provide feedback regarding its performance under field-testing conditions. Raters (N = 13) applied the rubric to 12 essays. Fit statistics indicate the raters are inconsistent in their use of the rubric. Reasons for this are discussed and suggestions for the further development of a consistently applicable rubric are noted.

Keywords

rubric, writing, FACETS

Introduction

Rubrics for evaluating compositions have existed for many years. One rubric that is well-known in Japan — the ESL Composition Profile (Hughey, Wormuth, Hartfiel, & Jacobs, 1983) — has been used for more than 25 years by many composition teachers. However, the number of items on this rubric means that it takes a significant amount of time to evaluate one paper. In addition, this rubric was designed for an ESL rather than an EFL environment with a specific first language.

Rubric use has increased as writing assessment has moved from indirect measures of writing ability to direct assessment (Slomp, 2008). In an overview on the use of rubrics in a variety of learning and assessment situations, Jonsson and Svingby (2007) make several conclusions regarding their use. Rubrics can allow for reliable scoring when they are topic specific and have potential for improving instruction. However, Jonsson and Svingby also indicate that rubrics do not facilitate valid judgment of the writing performance but suggest that this can be addressed "by using a more comprehensive framework of validity when validating the rubric" (p.141). One issue in rubric use is the influence of the rubric on the classroom, as the design is likely to either lead to a narrowing of the educational focus as teachers replace their rubrics with standardized

＊ヴィスゲイティス　ブラッド：大阪国際大学人間科学部教授〈2010.9.29受理〉

ones (Mabry, 1999), or, as Crawford & Smolkowski (2008) indicate, the design of the assessment instrument influences the pedagogical focus, a concern that Slomp (2008) shares.

In 2009, a Kansai Regional working group for JACET (Japan Association for College English Teachers) applied for and received special research funding from the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) to develop a rubric for evaluating student compositions that would have both good validity and reliability while at the same time being simple to use. The result of their initial efforts is a rubric (see Appendix A) that consists of five broad "dimensions," each of which encompasses a number of more specific elements. According to the working group's instructions, each dimension is to be rated holistically, using a point scale decided by the teacher (though a 5-point scale is provided on the rubric by the working group), and each dimension is to receive the same weight (e.g., each dimension is worth the same maximum number of points).

Discussions with one of the members of the JACET working group indicates that the group wants to develop a rubric that will yield consistent results on evaluations in a variety of educational settings (including both program- and class-level), with a wide assortment of teachers, and for a broad selection of essays. The purpose of this paper is to use Rasch analyses to provide feedback to the working group concerning the performance of the rubric under one field-testing condition.

The Study

Data Collection

Data collection began in early December 2009 when a number of teachers were contacted and asked to participate. Interested teachers (raters) were given a short explanation sheet for the project, a sheet that explained the rubric, and a set of essays with the scoring rubric (see Appendix A). The raters were permitted to participate to any degree they saw fit. As a result, not every rater chose to read or evaluate all of the essays (see Table 1).

Materials

Twelve 1- to 2-page essays (see Table 2) that had been collected during a regular composition course were used as the materials. The prompt for the essays was "What are the advantages and disadvantages of the place you are living?" However, students were allowed great latitude in their interpretation of the topic, resulting in a number of different themes. The essays ranged from 336 to 681 words in length, from 6 to 11 paragraphs, from 56.7 to 87.0 points in Flesch Reading Ease, and from 3.5 to 8.6 in grade levels on the Flesch-Kincade grade level of readability.

Table 1. Matrix of Overall Ratings for Essays by Rater

		Essays													
Raters	L	1	2	3	4	5	6	7	8	9	10	11	12	*M*	*SD*
1	E	18	20	16	14	20	20	18	15	17	21	19	13	17.58	2.61
2	E	19	22	18	18	22	20	20	20	21	19	20	22	20.08	1.44
3	E	17	17	18	18	19	16	19	13	12	16	17	15	16.42	2.19
4	E	15	16	17	18	18	15	15	14	16	19	18	20	16.75	1.86
5	J	15	25	12	15	13	19	10	14	12	12	12	23	15.17	4.73
6	J	17	21	16	12	16	18	13	11	14	17	15	14	15.33	2.77
7	J	11	18	20	12									15.25	4.43
8	E		19		20		18			18			19	18.80	0.84
9	J									14	17	14	14	14.75	1.50
10	E	16	23	19	19	21	17	19	16	17	24	22	25	19.83	3.13
11	E	16	25	17	20	17	19	19	13	21	18	22	21	19.00	3.13
12	J					16	13	9	13					12.75	2.87
13	E	17	18	11	14	21	13	15	12	14	12	15	12	14.50	2.94
	M	16.10	20.36	16.40	16.36	18.30	17.09	15.70	14.10	16.00	17.50	17.40	18.00		
	SD	2.18	3.11	2.88	3.04	2.83	2.55	3.97	2.51	3.16	3.69	3.41	4.54		
Length		427	576	333	337	678	670	482	443	621	610	633	485		

Note: E = Native English Speaker, J = Native Japanese Speaker

Table 2. Essay Characteristics

			Average			Readability		
Essay	Words	Paragraphs	S per P	W per S	C per W	Pass %	FRE	F-K
1	427	6	8.8	9.7	3.9	2	87.0	3.5
2	579	11	5.3	12.0	4.6	4	60.1	7.8
3	336	8	5.8	11.4	4.6	3	64.3	7.1
4	340	8	5.6	12.0	4.2	0	67.9	6.7
5	681	8	10.0	13.4	4.6	2	56.7	8.6
6	673	9	8.3	13.2	4.0	4	79.4	5.4
7	485	8	7.8	12.3	4.5	2	60.5	7.8
8	446	8	6.3	11.7	4.5	0	62.2	7.4
9	624	11	5.6	12.4	4.6	0	61.6	7.7
10	615	8	8.8	13.8	4.8	0	57.5	8.5
11	636	11	9.9	7.9	4.5	3	71.8	5.2
12	485	11	4.9	12.2	4.7	2	61.7	7.6
M	527	9	7.3	11.8	4.5	2	65.9	6.9
SD	124	2	1.9	1.6	0.3	2	9.2	1.5
Min	336	6	4.9	7.9	4	0	56.7	3.5
Max	681	11	10.0	13.8	5	4	87.0	8.6

Notes: S per P = Sentences per paragraph, W per S = Words per sentence, C per W = Characters per word, Pass % = Percentage of passive sentences, FRE = Flesch Reading Ease, F-K = Flesch-Kincaid Reading Level

Raters

Raters consisted of 13 experienced EFL writing class teachers from two universities in the Kansai area of Japan. Eight of the raters were native English speakers. Five

of the raters were native Japanese speakers with near-fluency in English. Although no specific data were collected concerning the raters' number of years of teaching experience or type of training, many of the raters have been known to the researcher for more than 10 years.

Analysis

A Rasch analysis using Facets 3.66 (Linacre, 2009) was performed with three facets, Essays, Raters, and Dimensions, which interacted to produce the measure, Quality. Appendix B shows the file specifications for the analysis. No problems were encountered in the analysis and subset connection was achieved in 31 iterations. Nevertheless, due to a very limited amount of data, these results must be considered to be highly tentative. More essays, more raters, and more evaluations need to be collected for a full analysis of the rubric.

Results

Essays

The essays were very similar in quality (see Table 3), with measures ranging from -1.02 logits (lowest quality) to 1.22 logits (highest quality), or slightly over 2 *SD*. Model population separation reached 2.71, with reliability estimate of .88. Model population

Table 3. Essay Measurement Report (arranged by mN).

Total Score	Total Count	Observed Average	Fair-M Average	Measure	Model S. E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrim	Corr. PtBis	# Essay
224	55	4.07	4.07	1.22	0.19	1.11	0.66	1.22	1.05	0.74	0.25	02
183	50	3.66	3.66	0.54	0.19	1.16	0.85	1.13	0.71	0.82	0.38	05
198	55	3.60	3.53	0.33	0.18	1.71	3.32	1.67	3.13	0.15	0.35	12
193	55	3.51	3.46	0.21	0.18	0.81	-1.10	0.82	-1.00	1.19	0.42	06
175	50	3.50	3.44	0.19	0.19	0.83	-0.90	0.82	-0.97	1.25	0.48	10
174	50	3.48	3.42	0.15	0.19	0.72	-1.62	0.72	-1.56	1.35	0.49	11
164	50	3.28	3.19	-0.22	0.19	0.83	-0.87	0.87	-0.64	1.13	0.43	03
180	55	3.27	3.16	-0.27	0.18	0.78	-1.26	0.77	-1.33	1.30	0.49	04
161	50	3.22	3.12	-0.33	0.19	0.88	-0.58	0.88	-0.61	1.13	0.46	01
176	55	3.20	3.08	-0.39	0.18	1.06	0.39	1.03	0.22	0.93	0.42	09
157	50	3.14	3.07	-0.41	0.19	1.08	0.45	1.06	0.39	1.00	0.49	07
141	50	2.82	2.72	-1.02	0.20	0.95	-0.21	0.92	-0.32	1.12	0.50	08
177.17	52.08	3.40	3.33	0.00	0.19	0.99	-0.07	0.99	-0.08		0.43	Mean
20.59	2.47	0.30	0.33	0.55	0.01	0.26	1.29	0.25	1.24		0.07	StDev (Pop)
21.50	2.57	0.32	0.35	0.57	0.01	0.27	1.34	0.26	1.30		0.07	StDev (Sample)

Model, Population: RMSE .19 Adj (True) S.D. .51 Separation 2.71 Strata 3.95 Reliability .88
Model, Sample: RMSE .19 Adj (True) S.D. .54 Separation 2.85 Strata 4.13 Reliability .89
Model, Fixed (all same) chi-square: 96.0 d.f.: 11 significance (probability): .00
Model, Random (normal) chi-square: 9.9 d.f.: 10 significance (probability): .45
Notes: Shading indicates misfitting items.

separation is, "the number of statistically different levels of performance that can be distinguished in a normal distribution with the same 'true' S.D. as the current sample" (Linacre, 2009, p. 258). The reliability of .88 is moderately good.

Examination of infit mean square statistics indicated only one of the twelve essays (Essay 12) to lie outside of the recommended range 0.75 – 1.30 (Bond & Fox, 2007, pp. 238-239), with a score of 1.71. Scores exceeding 1.30 are considered to be too erratic. This suggests that raters are focusing on different essay attributes when assigning scores.

Raters

Rater severity exhibited a similarly narrow range of scores (see Table 4), from -2.06 (least severe) to 0.51 (most severe). Here, the model population separation index was 3.35 with a reliability of .92, both acceptable figures.

Greater agreement between raters and less of a spread in the scores that are assigned would be a strong indication that the rubric was being followed. However, even if there is a wide variation in scores, systemic variation, such as when one rater is consistently severe, can be managed by systemically adjusting their scores (e.g., Bond

Table 4. Rater Measurement Report (arranged by mN).

Total Score	Total Count	Obsvd Average	Fair-M Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrim	Corr. PtBis	Exact Obs%	Exact Exp%	# Rater
51	20	2.55	2.58	0.51	0.32	1.19	0.66	1.15	0.54	0.79	0.21	28.1	28.2	12
174	60	2.90	2.86	-0.02	0.18	1.45	2.30	1.39	2.01	0.64	0.46	28.7	31.6	13
59	20	2.95	2.89	-0.06	0.30	0.91	-0.20	0.91	-0.20	1.01	0.01	31.6	31.2	9
61	20	3.05	2.97	-0.20	0.30	1.59	1.75	1.65	1.89	0.25	0.20	23.2	33.1	7
182	60	3.03	3.01	-0.27	0.18	1.62	3.05	1.52	2.63	0.52	0.40	29	33.1	5
184	60	3.07	3.05	-0.33	0.17	0.92	-0.42	0.90	-0.51	1.02	0.46	34	33.4	6
197	60	3.28	3.28	-0.72	0.17	1.21	1.23	1.39	2.08	0.63	0.09	32.7	34.8	3
206	60	3.43	3.45	-0.99	0.17	0.72	-1.79	0.75	-1.57	1.40	0.53	37.7	35	4
211	60	3.52	3.54	-1.13	0.17	0.61	-2.64	0.63	-2.44	1.37	0.39	35.2	34.9	1
94	25	3.76	3.68	-1.36	0.27	0.56	-1.90	0.70	-1.14	1.34	0.34	38.8	35.3	8
228	60	3.80	3.86	-1.64	0.18	1.01	0.14	1.00	0.03	1.01	0.40	31.5	33.1	11
238	60	3.97	4.04	-1.96	0.18	0.71	-1.76	0.68	-1.90	1.43	0.50	35.4	31.2	10
241	60	4.02	4.09	-2.06	0.18	0.64	-2.30	0.60	-2.45	1.50	0.49	36.8	30.6	2
163.54	48.08	3.33	3.33	-0.79	0.21	1.01	-0.14	1.02	-0.08	0.99	0.34			Mean
68.32	17.92	0.44	0.47	0.78	0.06	0.36	1.79	0.35	1.73	0.39	0.16			StDev(Pop)
71.11	18.66	0.46	0.49	0.81	0.06	0.37	1.87	0.36	1.80	0.40	0.17			StDev(Sample)

Model, Population: RMSE .22 Adj (True) S.D. .74 Separation 3.35 Strata 4.80 Reliability .92
Model, Sample: RMSE .22 Adj (True) S.D. .78 Separation 3.50 Strata 4.99 Reliability .92
Model, Fixed (all same) chi-square: 179.0 d.f.: 12 significance (probability): .00
Model, Random (normal) chi-square: 11.1 d.f.: 11 significance (probability): .43
Inter-Rater agreement opportunities: 2950 Exact agreements: 977 = 33.1% Expected: 972.4 = 33.0%
Shading indicates misfitting items.

& Fox, 2007, p. 157).

However, infit mean square statistics indicated that responses by eight of the 13 raters did not have good fit to the model. Five of the raters had scores below 0.75 (see Table 4). This indicates much less variation in their ratings than expected by the model, which might be a symptom of using too few rating categories for each of the dimensions. These muted responses may be due to lack of interest on the part of the raters (response set), a poor conceptualization of the differences between the five levels (which are not explicitly defined) for each dimension in the rubric, lack of adherence to the rubric when assigning scores, or some combination of these factors. In contrast, there were another three raters whose infit mean square statistics were in excess of 1.30. This indicates greater inconsistency in scoring. This also suggests the need for rater data to be further examined along the same lines as the examination of the essays. The lack of good fit to the model may be due to the same set of factors. Both of these issues will need to be addressed going forward. More detailed analyses should be done following the procedure outlined in the Facets manual (Linacre, 2009, p. 255).

One other statistic that is important is inter-rater agreement. Facets models raters to be "independent experts" rather than "scoring machines," and therefore, inter-rater agreement or inter-rater reliability is not expected to be overly high (Linacre, 2009, pp. 162-163). Statistics for inter-rater agreement (see Table 4) show that of 2,950 opportunities for agreement, there were only 977 exact agreements, or 33.1%. This conforms well to the expected value for the model: 972.4 (33.0%) exact agreements.

Facets is also able to provide detailed information about which particular responses were most unexpected. Table 5 lists the unexpected responses with standardized residuals in excess of absolute 2. Regarding the unexpected responses, a number of issues may be influencing the results. First some of the results were not unexpected. The four raters who only turned in a portion of the ratings were listed. Interestingly, the rater with the greatest number of unexpected responses (7) was one of the members of the working group that developed the rubric. It may be that familiarity with the rubric itself and how it was developed provided that rater a different understanding of its features.

Finally, in Rasch analyses it is also possible to examine in greater precision the bias and interaction between facets. Figures 1 through 4 show the bias/interaction between raters and essays. These provide visual insight into how measures differ between essays and raters. For example, Figure 1 illustrates quite clearly the misfit of Essay 12. Figure 2 displays the bias interaction between essay and rater, showing similar misfit for Essay 12. Figures 3 and 4 provide this information for essay results relative to the overall measure.

Table 5. Unexpected Responses in order of Absolute Value of Standard Residuals

Cat	Score	Expected	Resd	StRes	Essay	Rater	Dimension
2	2	4.4	-2.4	-3.6	Essay 12	3	5, Mechanics
3	3	4.7	-1.7	-3.3	Essay 02	3	5, Mechanics
2	2	4.1	-2.1	-2.8	Essay 11	5	5, Mechanics
2	2	4.1	-2.1	-2.8	Essay 12	9	5, Mechanics
1	1	3.2	-2.2	-2.8	Essay 12	6	2, Organization
2	2	4	-2	-2.7	Essay 09	3	5, Mechanics
1	1	3	-2	-2.6	Essay 12	13	1, Content and Idea Development
2	2	3.8	-1.8	-2.4	Essay 02	4	4, Vocabulary
4	4	2.4	1.6	2.4	Essay 03	7	3, Grammar
5	5	3.1	1.9	2.4	Essay 05	13	1, Content and Idea Development
5	5	3.1	1.9	2.4	Essay 05	13	2, Organization
4	4	2.4	1.6	2.4	Essay 07	13	4, Vocabulary
1	1	2.8	-1.8	-2.4	Essay 09	6	2, Organization
5	5	3.1	1.9	2.4	Essay 12	5	1, Content and Idea Development
5	5	3.2	1.8	2.3	Essay 02	5	3, Grammar
2	2	3.8	-1.8	-2.3	Essay 04	7	5, Mechanics
5	5	3.2	1.8	2.3	Essay 08	13	5, Mechanics
2	2	3.8	-1.8	-2.3	Essay 09	5	5, Mechanics
5	5	3.2	1.8	2.3	Essay 12	5	2, Organization
2	2	3.8	-1.8	-2.2	Essay 07	5	5, Mechanics
2	2	3.7	-1.7	-2.2	Essay 12	1	2, Organization
4	4	4.8	-0.8	-2.1	Essay 02	8	5, Mechanics
4	4	2.5	1.5	2.1	Essay 03	7	4, Vocabulary
1	1	2.4	-1.4	-2.1	Essay 03	13	4, Vocabulary
2	2	3.7	-1.7	-2.1	Essay 03	11	2, Organization
2	2	3.7	-1.7	-2.1	Essay 05	11	3, Grammar
2	2	3.7	-1.7	-2.1	Essay 06	12	5, Mechanics
5	5	3.4	1.6	2.1	Essay 08	5	5, Mechanics
2	2	3.7	-1.7	-2.1	Essay 12	1	1, Content and Idea Development
1	1	2.3	-1.3	-2	Essay 01	7	3, Grammar
5	5	3.4	1.6	2	Essay 02	5	4, Vocabulary
1	1	2.3	-1.3	-2	Essay 07	12	1, Content and Idea Development

Dimensions

Measure scores for the dimensions (see Table 6) showed the widest variation, ranging from -1.61 (least severe) for the dimension of mechanics to 0.86 (most severe) for grammar. In general, teachers were less forgiving of grammatical errors in essays than in other dimensions. The model population separation was 6.94, with a reliability .98.

Mechanics was the only dimension to exhibit poor fit, with an infit mean square score of 1.57. Again, scores exceeding 1.3 can be characterized as noisy and may be an indication that the raters are interpreting the scoring rubric in significantly different or inconsistent ways. This is clearly reflected in Table 5, where 12 of the 32 unexpected responses involved the dimension of Mechanics.

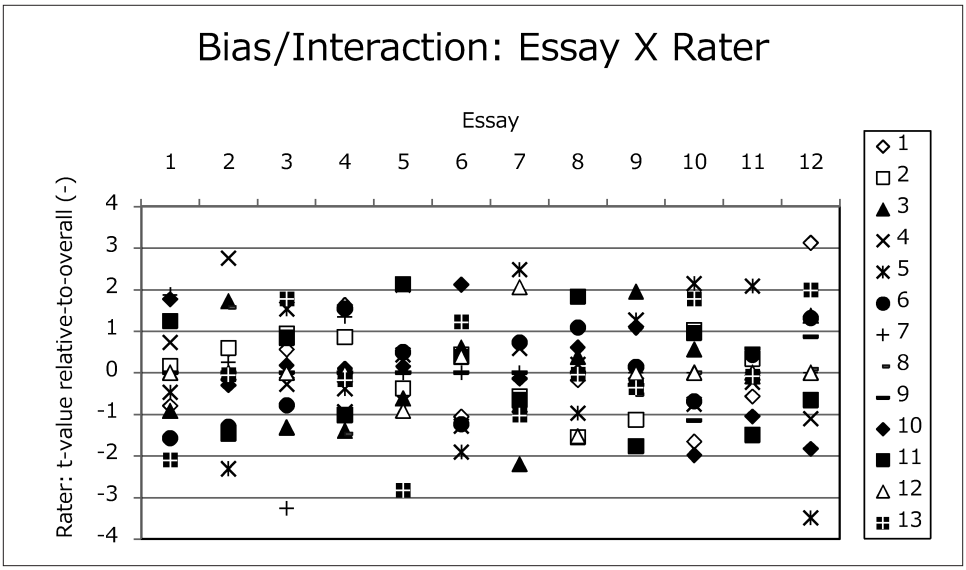

Figure 1. Bias/Interaction between rater and essay (T-value relative to overall.)

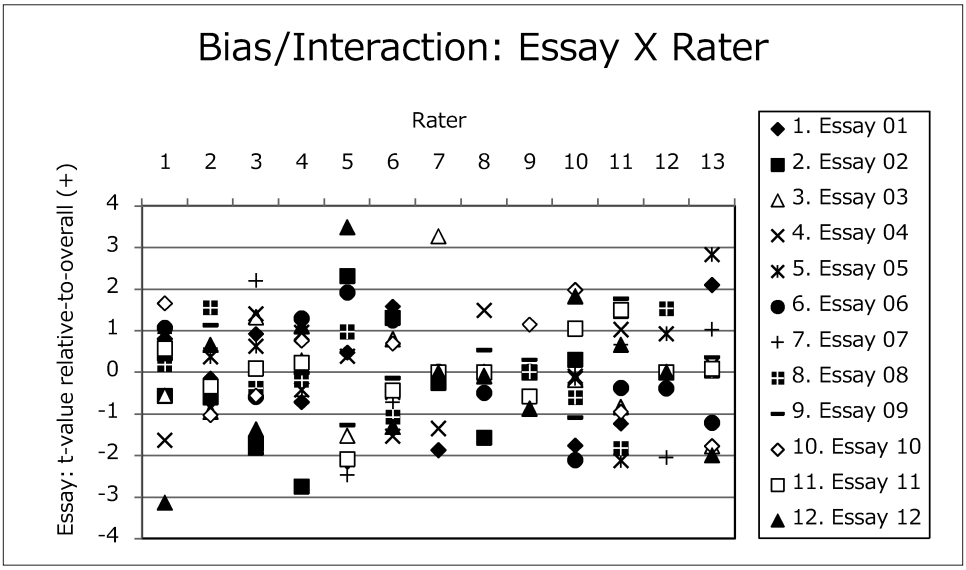

Figure 2. Bias/Interaction between essay and rater (T-Value relative to overall).

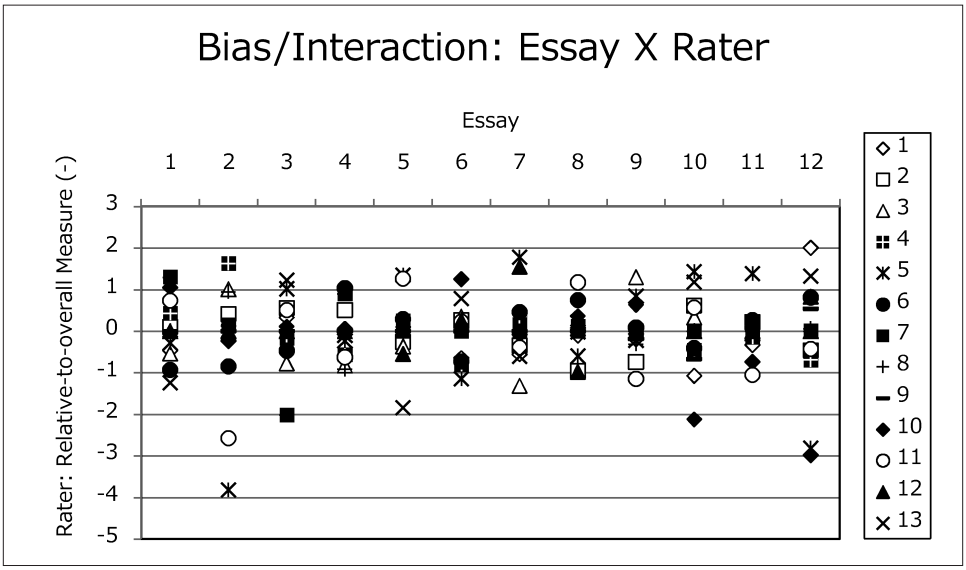

Figure 3. Bias/Interaction between rater and essay (relative to overall measure).

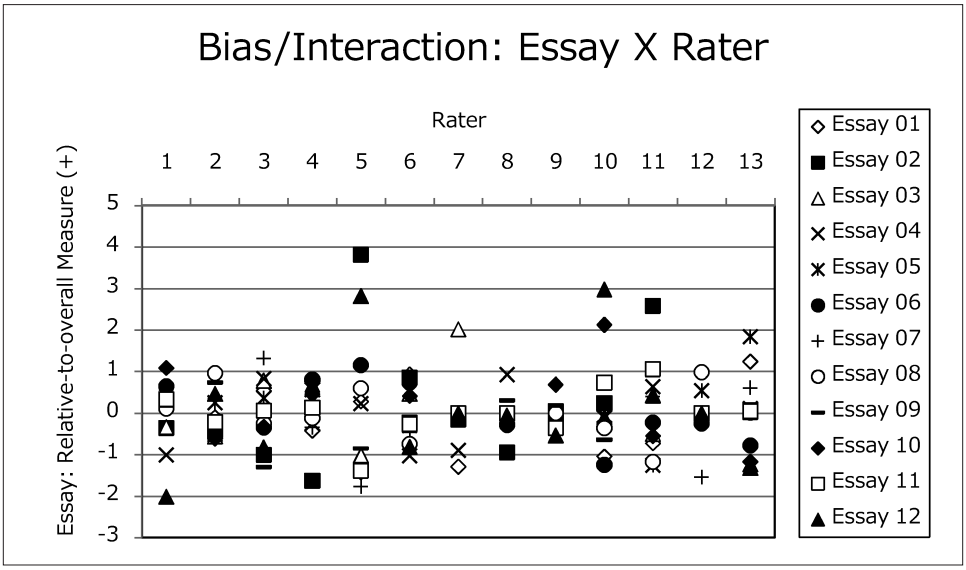

Figure 4. Bias/Interaction between essay and rater (relative to overall measure).

Table 6. Dimension Measurement Report (arranged by mN).

Total Score	Count	Observed Average	Fair-M Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrim	Corr. PtBis	# Dimension
366	125	2.93	2.81	0.86	0.12	0.76	-2.14	0.77	-2.00	1.23	0.39	3, Grammar
386	125	3.09	2.98	0.57	0.12	0.81	-1.66	0.82	-1.56	1.21	0.39	4, Vocabulary
419	125	3.35	3.26	0.11	0.12	0.89	-0.92	0.90	-0.87	1.11	0.41	1, Content and Idea Development
422	125	3.38	3.28	0.07	0.12	1.07	0.61	1.05	0.45	0.92	0.39	2, Organization
533	125	4.26	4.27	-1.61	0.14	1.57	3.88	1.45	2.94	0.48	0.24	5, Mechanics
425.20	125	3.40	3.32	0.00	0.12	1.02	-0.05	1.00	-0.21	0.99	0.36	Mean
57.81	0	0.46	0.51	0.86	0.01	0.29	2.17	0.25	1.78	0.28	0.06	StDev (Pop)
64.64	0	0.51	0.57	0.96	0.01	0.33	2.43	0.27	1.99	0.31	0.07	StDev (Sample)

Model, Population: RMSE .12 Adj (True) S.D. .85 Separation 6.94 Strata 9.58 Reliability .98
Model, Sample: RMSE .12 Adj (True) S.D. .95 Separation 7.77 Strata 10.69 Reliability .98
Model, Fixed (all same) chi-square: 215.4 d.f.: 4 significance (probability): .00
Model, Random (normal) chi-square: 3.9 d.f.: 3 significance (probability): .27
Shading indicates misfitting item

Variation

Ideally, those traits you are targeting through an evaluation rubric (e.g., the five dimensions) should account for the main portion of the score variance. Unfortunately, using this rubric in combination with these particular essays and raters resulted in only slightly more than 47% of variance (see Table 7) being accounted for by the measure. More than 52% of the variance was accounted for by other factors, including just over 21% of which was caused by bias interaction between raters and essays.

Yardstick

An illustration of the levels in logits for the elements within the three facets, Essays, Raters, and Dimensions, is shown in Figure 5. This type of yardstick makes it possible to compare essays, raters, and dimensions on the same scale.

Category and Model Functioning

Each of the five dimensions was to be rated on a 5-point scale, with each of the dimensions receiving equal weighting. Table 8 shows the percentages of each of the rating categories. Ratings were negatively skewed, meaning more ratings of 4s and 5s were assigned than 1s and 2s. Category probability curves are shown in Figure 6. Ogives (using Model = ?B, ?B, ?, Quality) that display category functioning relative to item difficulty are shown in Figures 6 to 9. Figure 6 provides the probability curve for rating category. Figure 7 gives the category information in measure relative to the item difficulty. Figure 8 illustrates the conditional probabilities. Figure 9 illustrates the cumulative probabilities. The roughly equal spacing and shape of the ogives indicates relative good category functioning. The degree to which the data fit the Rasch model

Table 7. Bias/Interaction: 1. Essay, 2. Rater

Raw-score variance of observations	=	100.00%
Variance explained by Rasch measures	=	47.09%
Variance of residuals	=	52.91%
Variance explained by bias/interactions	=	21.41%
Variance remaining in residuals	=	31.50%

Table 8. Category Statistics

Data				Quality Control			Rasch-Andrich Thresholds		Expectation Measure at		Most probably from	Rasch Thurstone Thresholds	Cat Peak Prob	Category Response Name
Category Counts Cum.				Avg.	Exp.	Outfit								
Score	Used	%	%	Meas	Meas	MnSq	Measure	S. E.	Category	-0.5				
1	8	1%	1%	-0.69	-0.77	1.0			(-4.28)		Low	low	100%	lowest
2	118	19%	20%	-0.17	-0.19	1.1	-3.17	0.36	-1.84	-3.31	-3.17	-3.23	66%	
3	213	34%	54%	0.49	0.51	1.0	-0.44	0.12	0.26	-0.66	-0.44	-0.55	50%	middle
4	187	30%	84%	1.38	1.39	0.9	1.07	0.10	1.9	1.07	1.06	1.06	50%	
5	99	16%	100%	2.47	2.43	1.0	2.54	0.13	(3.77)	2.93	2.54	2.71	100%	highest
									(Mean)		(Modal)	(Median)		

is represented in Figure 10. In general, the empirical ICC runs relatively near to the expected score ogive, and with only a few data point outside the 95% confidence interval.

Discussion

There are a number of problems with this project. First, fit statistics indicate that the raters are not scoring essays with the necessary consistency. This may be attributable to three interacting factors: insufficient training of the raters in the use of the rubric, insufficient specificity within the rubric (in both the descriptions of the dimensions that are targeted and the distinct levels within each dimension), and lack of comprehensiveness in the rubric itself (leaving other salient features out of the rating).

Going Forward

In the near term, there are a number of further statistical analyses that could shed more light on how the rubric and raters are functioning. One strategy would be to expand the model to include other variables, such as the essay characteristics listed in Table 2. These could easily be incorporated into DIF analyses, either directly or through the use of dummy variables. A second strategy would be to examine rater behavior more closely. One possibility would be to incorporate other rater factors (age, experience, training, first language) into the analyses and estimate to what extent those factors might influence the results. Finally, more data clearly needs to be collected in order to make any statistical analyses more robust.

Regarding the essays, follow-up analyses need to be conducted on Essay 12 in order

Measure	Essay (Better)	Rater (Severe)	Dimension (Severe)	Quality (High)
2				(5)
				4
	2			

1				
			Grammar	
			Vocabulary	
	5	12		
	12			3
	6 10 11			
			Content and Idea Development Organization	
0		13		
		9		
	3	7		
	1 4	5 6		
	7 9			
		3		---
-1	8	4		
		1		
		8		
		11	Mechanics	
				2
-2		10		
		2		
-3				(1)
Measure	Essay (Worse)	Rater (Lenient)	Dimension (Severe)	Quality (Low)

Figure 5. Yardstick of measures for Essays, Raters, and Dimensions.

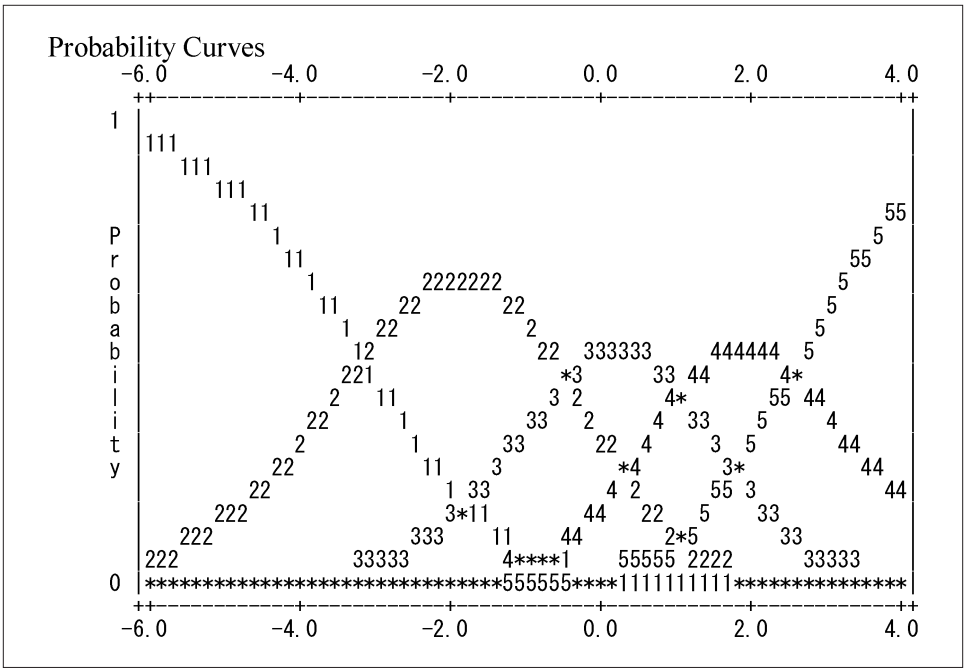

Figure 6. Probability curves for rating categories.

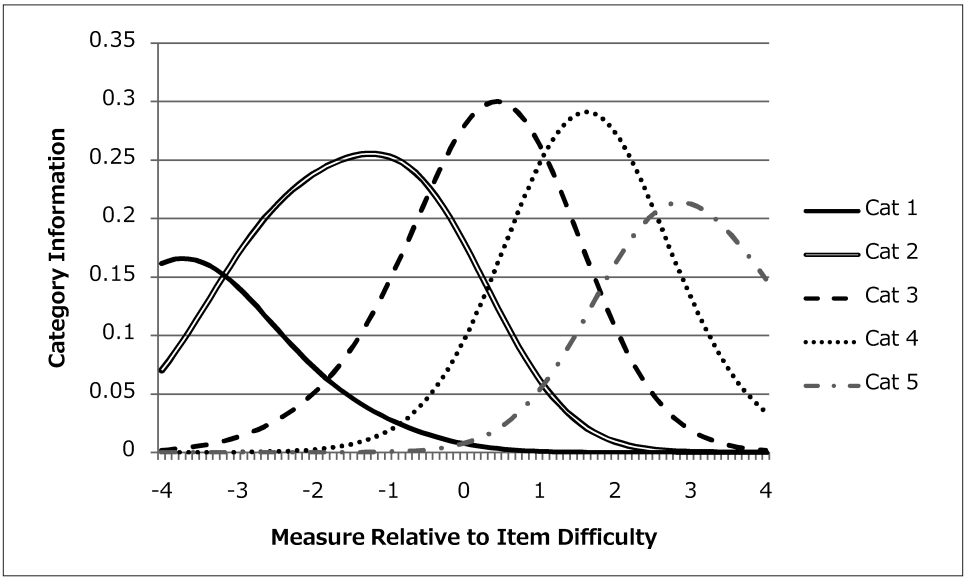

Figure 7. Category information in measure relative to item difficulty.

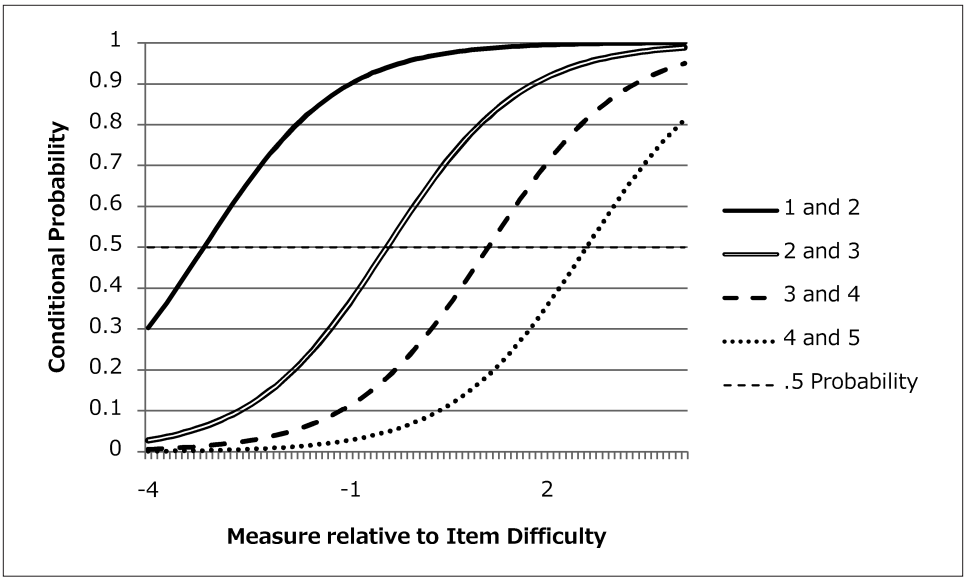

Figure 8. Conditional Probabilities in Measures Relative to Item Difficulty

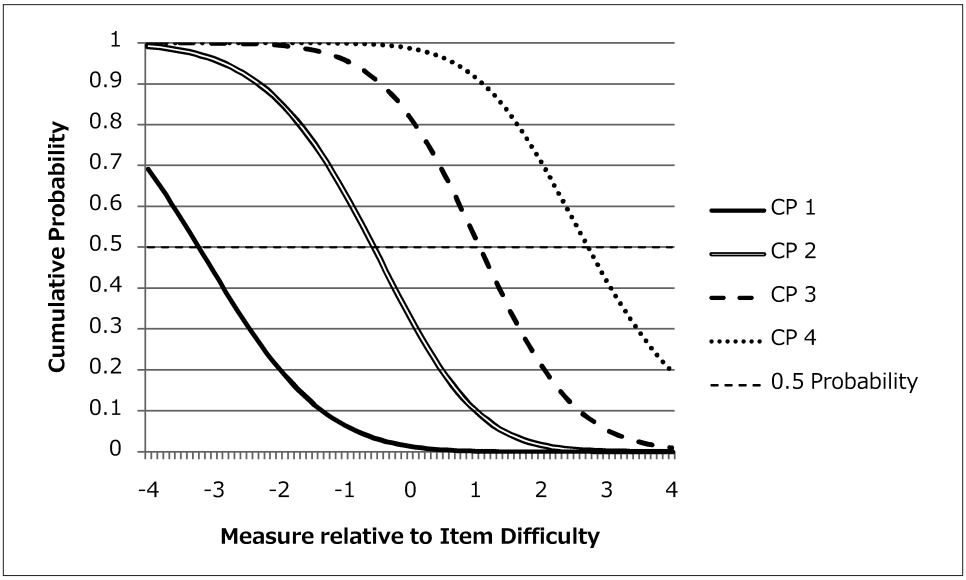

Figure 9. Cumulative probabilities in measures relative to item difficulty.

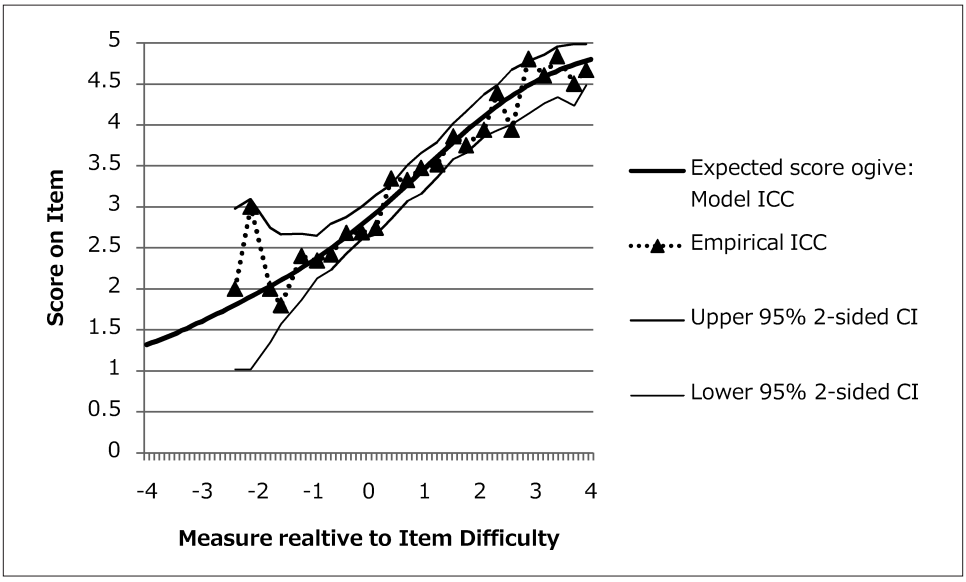

Figure 10. Expected score Ogive, Empirical ICC and 95% CI.

to determine possible causes for the misfit, including such aspects as examination of rater-essay bias, L1 of the rater, and the influence of other essay characteristics (see Table 2).

In the mid term, raters who showed the largest number of unexpected responses could be targeted for "mini case studies" that could address issues involving interpretation of the rubric, determination of quality level within each of the dimensions, and other salient features within the essays that seemed to influence scoring decisions. These could then be incorporated into future rubrics.

Another mid-term step could be to improve the way in which raters are guided in the application of the rubric. For example, more explicit instruction as to how to distinguish between levels within a dimension would most likely result in more consistent ratings. So, too, would provision of some previously marked exemplars accompanied by short explanations of the ratings. Both of these steps would make application of the rubric, at least initially, more time consuming, but likely result in more accurate and consistent results. As Bachman's (2000) review of language assessment indicates, a number of studies on ESL student writing have found differences in rater behavior on the ways in which assessments are made, clearly suggesting training to be a feature that needs to be considered with the design of any rubric.

Over the longer term it would probably be more beneficial for the project team to follow procedures outlined in Wilson (2005) concerning the construction of measures.

This would involve more carefully defining the targeted construct (What is good EFL composition?), identifying the dimensions that contribute to that construct (type) as well as the degree to which they contribute (weight), and then determining the precise characteristics within each dimension that would distinguish between an L2 writer who demonstrated greater ability from an L2 writer who demonstrated lesser ability for that particular skill. Ideally, these should form a type of Guttman stepping sequence, where an individual at a particular level of competence would in that scale also be competent at all of the levels below that one. If that level of precision is not possible, then a better idea of the aims of the rubric itself need to be clarified. In any case, if the ultimate goal is development of an invariant measure, then much more work needs to be done to meet the five necessary criteria outlined by Engelhard (December 18-20, 2009, course handouts).

Huang (2008) discusses the problems with the rating of ESL students' writing in the second-language environment, such as would be found in the U.S., and indicates that there was greater inconsistency on the assessment of ESL students' writing than for that by native English (NE) writers. Huang makes four specific recommendations for assessment of ESL writing: adjudication to reduce discrepancies in assessments, greater attention to task design, training specifically on ESL writing assessment in addition to that provided for NE writing, and better tracking of individual raters. Familiarity with the criteria is recommended in a number of evaluations of writing assessment (e.g., Martin & Penrod, 2006; Penny, Johnson, & Gordon, 2000; Slomp, 2008). Rubric familiarity would seem to be a benefit for foreign language writing (EFL) evaluations as well. An additional issue with rubrics is the difference between the rubric itself and the actual second-language writing tasks students are expected to perform at the university level, as a standardized rubric may be inappropriate for a variety of academic writing tasks (Moore & Morton, 2005). Clearly, these are issues that the JACET working group should be encouraged to consider.

Finally, prior to these steps, it would probably be very useful to determine the precise purpose of the rubric. To my way of thinking, there are two possible but largely incompatible purposes. The first is for evaluating students, such as when assigning grades or doing placement. For this purpose, invariance would be the most important aspect, and only those dimensions that had bearing upon the grading or placement decision would need to be included. A second purpose is to generate feedback for diagnostic and pedagogical aims. With this purpose in mind, it would be better to have dimensions and levels within the dimensions that corresponded to areas where students were capable of improving their performance. These would need to be features that were both salient and malleable over the short term. In such pedagogical cases, inclusion of other elements and dimensions would not serve the teachers, nor

the students, very well. They would also need to be at very precise and narrow levels, thereby reducing their utility for placement or general evaluation purposes.

References

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Crawford, L., & Smolkowski, K. (2008). When a "sloppy copy" is good enough: Results of a state writing assessment. *Assessing Writing, 13*, 61-77.

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

Hughey, J. B., Wormuth, D. R., Hartfiel, V. F., & Jacobs, H. L. (1983). *Teaching ESL composition: Principles and techniques*. Rowley, MA: Newbury House.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*, 130-144.

Linacre, J. M. (2009). Facets Rasch measurement computer program, version 3.66.0 (Version 3.66.0). Chicago: Winsteps.com.

Mabry, L. (1999). Writing to the Rubric. *Phi Delta Kappan*, May, 673-679.

Martin, D., & Penrod, D. (2006). Coming to know criteria: The value of an evaluating writing course for undergraduates, *Assessing Writing, 11*, 66-73.

Moore, T., & Morton, J. (2005). Dimensions of difference: a comparison of university writing and IELTS writing. *Journal of English for Academic Purposes, 4*, 43-66.

Penny, J., Johnson, R., & Gordon, B. (2000). The effect of rating augmentation on inter-rater reliability An empirical study of a holistic rubric. *Assessing Writing, 7*, 143-164.

Slomp, D. H. (2008). Harming not helping: The impact of a Canadian standardized writing assessment, *Assessing Writing, 13*, 180-200.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Appendix A

JACET Working Group Rubric 2009

ルーブリック 2009 / Rubric 2009

Essay Code #:

Evaluator #:

内容・展開　Content & Idea Development =	1	2	3	4	5

- The content is relevant to the given topic 課題の要求に対応した内容が述べられている
- The writing was completed in accordance with the task requirements 課題の条件（字数・時間など）を満たして作文を完成している
- The writing is coherent 論旨から外れることなく，首尾一貫した主張を展開している
- Every topic sentence is clearly shown トピックセンテンス（トピック＋メインアイディア）は，明確に提示されている
- Every topic sentence is fully developed トピックセンテンス（トピック＋メインアイディア）は，十分に展開されている

構成　Organization =	1	2	3	4	5

- The writing includes an introduction – body - conclusion structure 導入・展開・結論の構成になっている
- The writing is logically organized 文章は論理的に構成されている
- The writing flows smoothly 文章の流れはスムーズである
- Connecting words / expressions are used effectively つなぎ言葉が効果的に使用されている

文法　Grammar	1	2	3	4	5

- A variety of sentence structures are appropriately used 多様な文構造が適切に使われている
- Sentence structures are accurate 文構造は正しい
- The writing has no errors of subject-verb agreement, tense, number, pronouns, articles, prepositions, and so on 主語・動詞の一致，時制，数，代名詞，冠詞，前置詞などに誤りがない

語彙　Vocabulary	1	2	3	4	5

- A variety of words and expressions are used 多様な語や表現が使われている
- The choices of words and expressions is appropriate 語や表現の選択は適切である

綴り・句読点など　Mechanics	1	2	3	4	5

- The first line of each paragraph is appropriately indented (3 to 5 letters or ½ inch) 段落の最初の行が適切に字下げ（3～5字または½インチ程度）されている
- The return key is only used between paragraphs (e.g., not at the end of each line or sentence) 段落の始めなど，適切な位置で改行されている
- The spelling is accurate 綴りが正確である
- The punctuation and capitalization are accurate 句読点・大文字の使用が正確である

Appendix B

File Specifications

```
Writing Rubric 2009 1/6/2010 11:48:07 AM
Table 1. Specifications from file "C:\Users\Visgatis\Desktop\Rubric\Rubric 3F.txt".

Title = Writing Rubric 2009 1/6/2010 11:48:07 AM
Data file = (C:\Users\Visgatis\Desktop\Rubric\Rubric 3F.txt)
Output file = C:\Users\Visgatis\Desktop\Rubric\Rubric 3F.out.txt

; Data specification
Facets = 3
Non-centered =  2
Positive =  1
Labels =
 1,Essay ; (elements = 12)
 2,Rater ; (elements = 13)
 3,Dimension ; (elements = 5)
Model = ?B,?B,?,QUALITY,1
Rating (or other) scale = QUALITY,R5,General,Ordinal

; Output description
Arrange tables in order = mN
Bias/Interaction direction = ability ; leniency, easiness: higher score = positive logit
Fair score = Mean
Pt-biserial = Yes
Heading lines in output data files = Y
Inter-rater coefficients reported for facet = 2
Omit unobserved elements = yes
Barchart = Yes
Total score for elements = Yes
T3onscreen show only one line on screen iteration report = Y
T4MAX maximum number of unexpected observations reported in Table 4 = 100
T8NBC show table 8 numbers-barcharts-curves = NBC
Unexpected observations reported if standardized residual >= 6
Usort unexpected observations sort order = u
Vertical ruler definitions = 1N,2A,3A
WHexact - Wilson-Hilferty standardization = Y

; Convergence control
Convergence = .5, .01
Iterations (maximum) = 0 ; unlimited
Xtreme scores adjusted by = .3, .5  ;(estimation, bias)
```