Psychometric Quality of the Writing & Speaking Subtests of LRN International English Language Competency Assessment (IELCA)

Prepared by
Yuyang Cai, The University Of Hong Kong

Index

1. Author biodata

Yuyang Cai is currently pursuing his PhD in language testing and assessment at Faculty of Education in the University of Hong Kong. He received his M.Ed. in English Language Education from Memorial University of Newfoundland in 2006 and his B.A. in English Language and Literature from Central China Normal University in 1996. His research interests include:

- language development and validation
- assessing English for Specific Purposes
- assessing reading
- assessing grammar
- language use strategy
- multidimensional item response theory
- cognitive diagnostic measurement
- structural equation modeling
- growth-curve modeling
- multi-level modelling

He is the chief designer of the Nationwide English for Specific Purposes Assessment Project in China and has had three years' experience in managing the development and validation of the project before his PhD study.

2. Introduction

The purpose of this report is to explore the psychometric quality of the writing & speaking subtests of LRN's International English Language Competency Assessment (IELCA) ratings using the many-facet Rasch measurement. In particular, it is to examine the extents to which rater severity affect IELCA writing and speaking subtests ratings and the psychometric dimensionality of the scoring criteria for IELCA writing and speaking subtests.

Language assessment programs are committed to gathering evidence for the validity of their assessments. A tough challenge in this process has been to develop a reliable scoring system, given the nature of the assessment and the complexity of the construct measured (Crocker, 1997). The ratings of writing and speaking test performance involve the nature of the rating scales and judge's interpretation and judgment on them with the risk of systematically recruiting rater effects and scale, and into examinees' scores being high, among others, thereby endangering the psychometric quality of the tests (Eckes, 2005). The former effects are commonly recognized as rater severity or leniency, halo, or central tendency, and are viewed as a source of systematic variance in observed ratings irrelevant to the examinees (Myford & Wolfe, 2003). This construct irrelevant variance, as suggested by research in L2 performance assessment, can be reduced (though not easily removed) through rater training (Weigle, 1998). The latter relates to the dimensionality of the scales or criteria domains, which concerns whether different criterion relates to one dimension or several different dimensions.

3. IELCA Writing and Speaking Subtests

The dataset analyzed in this report was from a trial administration of the International English Language Competency Test (IELCA) in 2013 which has two versions: one for academic purpose (abbreviated as AC) and one for general training purpose (abbreviated as GT). IELCA writing (both AC and GT versions) is designed to measure the examinee's ability to respond appropriately to a given prompt, to organize a piece of writing, to use a range of lexical and grammatical items accurately, and to show awareness of audience and genre. The IELCA writing subtest consists of two tasks: the first being a prompt with text provided where the examinee is asked to respond in the format of a report (the AC version) or a letter/email (the GT version) and the second with the examinee being expected to produce an argumentative essay (the AC version) or longer discussion scripts (the GT version). Both the AC and GT version had two groups of examinees, which were labeled as AC Task1, AC Task2, AC Task-1, AC Task-2, GT Task 1, GT Task 2, GT Task-1, GT Task-2, respectively. Specific criteria for scoring the examinee performances are

specified in the rubric.

LRN's IELCA speaking subtest consists of three tasks: the first task is designed to measure the examinee's ability to answer questions relating to their personal life and general introductory questions involving a quick transaction between the examiner and candidate; the second is designed to measure the examinee's ability to express opinions on a given topic through the long turn (in an extended utterance); and the third task aims to measure the examinee's ability to enter a final transaction, based on the topic covered in section 2, at a deeper level. LRN's IELCA speaking subtest has also had two groups of examinees. According to tasks taken, the six datasets were labeled as Speaking Task 1 to Task 3, and Speaking Task-1 to Task-3. As was with the writing subtest, aspects to be evaluated as scoring criteria are stated in the rubric.

There are two routes offered for the IELCA writing – IELCA Academic (AT) and IELCA General (GT). Both subsets consist of two tasks: the first task is designed to measure examinee's ability to organise a piece of writing that responds to a given prompt – for example, a letter (GT) or describing a process or reporting activity in a graph (AT); the second task is discursive and aims to measure the examinee's ability to respond appropriately to a given prompt. Aspects to be evaluated as scoring criteria are stated in the rubric.

4. Research Questions

The main research questions addressed in this report are:

1. Do IELCA raters differ in the rating severity when rating performance within IELCA writing and speaking tasks; and, if so, to which extent?

2. What is the psychometric dimension of IELCA writing and speaking subtests? Are the criteria clearly distinguishable?

5. Methods

Examinees

LRN's IELCA writing subtest (AC and GT) was administered to 600 participants. IELCA (AC) was administered to 300 participants (104 females, 196 males). Participants' mean age was 24.50 (SD=5.38) and 91.7% of the participants were aged between 17 and 32 years. A further 300 participants took the GT version (80 females and 220 males) with the mean of their ages reaching 24.35 (SD=4.86). 93.0% of the participants were aged between 17 and 32 years. All

AC and GT participants were from four test centers, one in each of the four countries (with percentages): Nigeria (33.8%), India (33.2%), Pakistan (19.2%) and Malaysia (13.8%).

LRN's IELCA speaking test was administered to 633 participants (193 females and 440 males). The participants' mean age was 24.45 (SD=5.07) with 92.6% of the participants being aged between 17 and 32 years. These participants were from four test centers, one in each of the four countries (with percentages): Nigeria (33.0%), India (31.3%), Pakistan (22.6%) and Malaysia (13.1%).

6. Raters

The raters who scored IELCA writing and speaking performance were all licensed by LRN upon the fulfillment of strict selection criteria. Before rating, they were trained and monitored so to be compliant with scoring guidelines. Six raters scored the IELCA writing test and six other raters provided scorings of the speaking test. The rating training proceeded in two sessions with the first session having the raters mark five writing scripts and five speaking assessments to practice the rating scale. Issues and problems surrounding the marking process such as the interpretation of levels, specifically the cut-off scores, were discussed by the lead rater(s) and trainers. Raters were then given a batch of five writing scripts and speaking assessments to rate off-site. In the second session, their rating performance (that is, how they had marked the writing scripts and speaking assessments) and their understanding of the rating scale was closely examined. The raters were reminded that it was natural to have a variety of severity levels but that they should attempt to maintain their severity throughout their marking tasks in addition to inconsistency in marking resulting in an invalidation of the scores.

7. Procedure

Participants were first presented with the writing section (30 minutes), followed by the speaking section (14 minutes). Both ratings of writing scripts and recorded oral responses to speaking were carried out based on a pre-established analytical rating scale used by IELCA raters. The scoring criteria for writing included a) task achievement, b) coherence and cohesion, c) lexical resources, and d) grammatical range and accuracy. The scale for Writing Task 1 (for both AC and GT versions) was 5, 10, 10, and 10 (with a total of 35 points) and 6, 12, 12, and 12 (with a total of 42 points) for the second. The scoring criteria for speaking included a) pronunciation, b) fluency, c) language accuracy and appropriacy, and d) task fulfillment. The scale for the first speaking task was 5 points for each criterion and that for the other two tasks 10 points.

In both the writing and speaking subtests, examinee responses were scored independently by four or two raters. These original scorings served as input to the Many-Facets Rasch Measurement analysis using FACETS Version 3.71 (Linacre, 2013a).

8. Data Analysis

All rating data were analyzed using the computer program FACETS, with separate FACETS analyses performed on each task of the writing and speaking subtests. The program used the ratings that raters awarded to examinees to estimate individual examinee proficiencies, rater severity, and criteria difficulties.

For IELCA writing, as the rating scale for each criterion of a task is unequal, the specific model implemented in the analyses was a three-facet partial credit model (Linacre & Wright, 2002). The estimation ceased automatically after 165, 127, 348, 581, 160, 114,110, 765 iterations for writing AC Task1, AC Task 2, AC Task-1, AC Task-2, GT Task1, GT Task 2, GT Task-1, GT Task-2, respectively, and after 93, 198, 153, 110, 143, 170 iterations for speaking Task 1, Task2, Task3, Task-1, Task-2 and Task-3, respectively.

9. Results

Global Model Fit

The FACETS calibrates the examinees, rater severity and rating criteria difficulty onto the same equal-interval scale (i.e., the logit scale) thus making the interpretation of analysis results possible by formulating a unified scale of reference. The overall data-model fit can be evaluated by examining unexpected responses given the assumptions of the model. The satisfactory model fit is indicated when about 5% or less of (absolute) standardized residuals are ≥ 2, and about 1% or less of (absolute) standardized residuals are ≥ 3(Linacre, 2013b).

Table 1 presents the global model fit statistics for all of the 14 tested models (i.e.,. 4 for AC writing, 4 for GT writing, and 6 for speaking subtests). Considering the writing tasks first - of the eight datasets analyzed, there were four with unexpected response percentage (absolute standardized residuals ≥ 2) larger than 5% : GT Task-1 (6.65%), GT Task 1 (5.83%), AC Task-2, and GT Task-2 (5.64%). Among them, only GT Task 1 was confirmed by examining the unexpected response percentage (1.32%, using the criterion of absolute standardized residuals ≥ 3). No other dataset showed a percentage larger than 1%. In terms of all speaking datasets, none of them displayed model-data misfit, applying either the cutoff criteria of absolute standardized residuals ≥ 2 or ≥ 3.

In all, these results showed a satisfactory model fit for most writing tasks (except for GT Task 1, about 7% of the number of tasks) and for all speaking subtest tasks. Other statistics (e.g., rater fit and criteria fit statistics) are provided later to further assess the dataset-model fit.

Table 1
Global Model Fit Results for IELCA Writing &Speaking Subtest

Part I: Writing								
	Academic (AC)				General Training (GT)			
	Task 1	Task 2	Task-1	Task-2	Task 1	Task 2	Task-1	Task-2
Responses with S.R. ≥2 (Percentage)	147 (4.59 %)	168 (5.28 %)	42 (5.25 %)	45 (5.65 %)	186 (5.83 %)	152 (4.80 %)	53 (6.65 %)	45 (5.64 %)
Responses with S.R. ≥3 (Percentage)	29 (.91%)	26 (0.82 %)	7 (0.88 %)	5 (0.63 %)	42 (1.32 %)	27 (0.85 %)	5 (0.63 %)	3 (0.38 %)
Total Valid Responses	3200	3184	800	796	3192	3166	797	798
Part II: Speaking								
	Academic (AC)				General Training (GT)			
	Task 1	Task 2	Task 3		Task 1	Task 2	Task3	
Responses with S.R. ≥2 (Percentage)	106 (3.31 %)	158 (4.94 %)	150 (4.69%)		128 (3.70 %)	161 (4.65 %)	152 (4.39%)	
Responses with S.R. ≥3	6 (0.19 %)	22 (0.69 %)	32 (1.00%)		9 (3.70 %)	9 (0.26 %)	31 (0.89%)	
Total Valid Responses	3200	3200	3200		3462	3464	3464	
Note: S.R. = standardized residual								

10. Calibration of writing and speaking subtests ratings

This section provides results of calibrations of writing and speaking subtests ratings (altogether 14 datasets). In practice, the many-facet Rasch analysis results are usually illustrated using the calibration map provided after running the FACETS program (see Appendices A to N for all 14 maps corresponding to each of the 14 datasets used in this study). A typical map for three facets usually consists of five or more columns. Take for example the IELCA writing AC Task 1 where the first column in the map displays this logit scale and the second shows estimates of examinee proficiency. The third column shows the severity variations among raters (from the most severe rater at the top and the least severe at the bottom) and the fourth compares the first criterion (task achievement) with the other three (i.e., coherence and cohesion, lexical resources, and grammatical range and accuracy) of IELCA AC writing scoring rubric in terms of their relative difficulties. Criteria located higher in the column were more difficult for examinees to receive high ratings than on criteria located lower in the column. Finally, the fifth column describes the five-point scale for the first criterion of task achievement and the last column describes the ten-point rating scales used for other three scoring criteria.

The effects of facets (i.e., rater severity and criteria difficulty) are evaluated by referring to statistics such as mean of measure (logit), standardised error of mean, chi-square (with degree of freedom and significance level), separation index, and separation reliability.

11. Rater effects

Question 1 in section 3 is addressed here which is also preceded by two sub-questions:

Question 1.1: Do IELCA raters differ in the rating severity when rating performance within IELCA writing tasks; and, if so, to which extent?

The variable maps illustrate the results of the calibrations of the examinees, raters and criteria for the eight writing datasets (see Appendices A to H, as corresponding to AC Task 1 to GT2). Table 2 gives different tasks labels, various statistics (i.e., rater labels, rater severity, error, and infit and outfit mean-square values) and other group statistics (mean, standard deviation of the mean, separation index fixed chi-square with degree of freedom, and significance level) accompanying each map.

Examining the variable maps will easily reveal that the rater severity measurements (levels) in each of the eight dataset are all located on the horizontal level with 0 logit value. This indicates that the severity level across

the four (or two) raters within each group were trivial or ignorable. With reference to Table 2, the largest severity span between the most lenient rater and the most severe rater was merely .06 logits in the AC Task 1 group. This was confirmed by the separation statistics, which all had the value of zeros. All fixed chi-square values with their corresponding degree of freedoms as well pointed to insignificant variation (with a smallest p value of .54) across rater severity levels. Therefore, the null hypothesis, that all raters were equally severe (lenient) within each group, cannot be rejected. These indicators of the magnitude of severity differences among raters indicate that significant variation in harshness did not exist among the raters. The fourth column in Table 2 shows that the level of error was small.

The fifth column presents two mean-square statistics indicating data-model fit for each rater: rater infit and rater outfit: the former is sensitive to an accumulation of unexpected ratings and the latter sensitive to individual unexpected ratings. Both of them can value from 0 to infinite, but with an expected value of 1(Linacre, 2002; Myford & Wolfe, 2003). Raters with fit values greater than 1 show more variation than expected and data provided by these raters tend to misfit the model. By contrast, raters with fit values less than 1 show less variation than expected in their ratings; data provided by these raters tend to overfit the model. As a rule of thumb, Linacre (2002) suggested to use .50 as a lower control limit and 1.50 as an upper control limit for infit and outfit mean-square statistics. Other researchers proposed a strict control of 0.70 (or 0.75) as lower limit and 1.30 as an upper limit (see Bond & Fox, 2001; McNamara, 1996).

Table 2
Raters Measurement Report for IELCA Writing Subtest

	Raters	Rater Severity Measure (in logits)	Error	Fit Infit MnSq	Fit Outfit MnSq	Group Statistics*
W-AC Task 1	R2 R1 R3 R4	-.02 -.01 .00 .04	.04 .04 .04 .04	.95 .74 1.18 1.16	.92 .74 1.13 1.12	M= .00, SD= .03, Separation =.00, fixed Chi-Square (df)= 1.4(3), *p*=.72.
W-AC Task 2	R1 R3 R2 R4	-.02 -.01 .00 .02	.03 .03 .03 .03	.89 .96 .98 1.02	.88 .99 .98 1.12	M=.00, SD=.01, Separation=.00, fixed Chi-Square (df)= .7(3), p=.88
W-AC Task-1	R2 R1	.00 .00	.07 .07	1.13 .76	1.15 .78	M=.00 , SD= .00, Separation=.00 ,fixed Chi-Square (df)=.00(1) , p= .97

W-AC Task-2	R2	.00	.04	.91	1.00	M=.00 , SD=.00 , Separation= .00, fixed Chi-Square (df)=4.99(1) , p=1.00.
	R1	.00	.04	.98	1.01	
W-GT Task 1	R3	-.01	.04	.97	.98	M=.00 , SD=.00 , Separation= .00, fixed Chi-Square (df)=.4 (3) , p=.93
	R1	.00	.04	.83	.80	
	R2	.00	.04	1.03	1.02	
	R4	.02	.04	1.08	1.13	
W-GT Task 2	R3	-.03	.04	.93	.94	M= .00, SD=.02 , Separation=.00 ,fixed Chi-Square (df)=2.22(3) , p=.54
	R1	-.02	.04	.97	.95	
	R2	.01	.04	.96	.97	
	R4	.03	.04	.95	1.05	
W-GT Task-1	R2	-.01	.05	.91	.87	M= .00, SD=.02, Separation=.00, fixed Chi-Square (df) =8 .99 (1), p=.84.
	R1	.00	.05	1.06	1.02	
W-GT Task-2	R2	.00	.04	.98	.98	M= .00, SD=.01, Separation=.00 , fixed Chi-Square (df)=5.19(1) , p=.97.
	R1	.00	.04	.94	.95	

Note: R1 to R4= Labels for Rater 1 to Rater 4, MnSq = mean square, M= Mean of Rater Logits per group, SD=standard deviation of rater logit per group, df= degree of freedom, p (value) =significance level, W-AC= writing for academic purpose, W-GT=writing for general training purpose

According to the fifth column, individual rater infit values ranged from .74 (by Rater 1 in AC Task 1) to 1.18 (by Rater 3 in AC Task 1) and the outfit values range from .74 (by Rater 1 in AC Task 1) to 1.15 (by Rater 2 in AC Task-1). If applying the wide range of lower and upper limits control, all rater severity levels are acceptable; while if applying the narrow range, Rater 1 slightly showed an overfit, which suggests a central tendency or halo effect (see Myford & Wolfe, 2004). However, given the small magnitude of .01 to .75, it can still be concluded that all these raters were internally consistent when rating IELCA writing tasks responses.

Question 1.2: Do IELCA raters differ in the rating severity when rating performance within IELCA speaking tasks; and, if so, to which extent?

The relationships among the examinees, raters and criteria for the six speaking datasets are mapped using the FACETS program (see Appendices I to N, as corresponding to Speaking Task 1 to Task-3). Table 3 gives different tasks labels, various statistics (i.e., rater labels, rater severity, error, and infit and outfit mean-square values) and other group statistics (mean, standard deviation of the mean, separation index fixed chi-square with a certain degree of freedom,

and significance level) accompanying each map.

Table 3
Raters Measurement Report for IELCA Speaking Subtest

	Raters	Rater Severity Measure (in logits)	Error	Fit		Separation Statistics*
				Infit MnSq	Outfit MnSq	
Speaking Task 1	R1 R2 R3 R4	-.15 -.04 .07 .12	.05 .05 .05 .05	.97 .99 1.00 1.05	.94 .97 .97 1.08	M= .00, SD= .10, Separation =1.82, fixed Chi-Square (df)= 17.2(3), p=.00.
Speaking Task 2	R1 R2 R3 R4	-.04 -.01 .00 .05	.05 .05 .05 .05	.82 .98 1.03 1.13	.82 .98 1.02 1.14	M=.00, SD=.03, Separation=.00, fixed Chi-Square (df)= 1.1(3), p=.63.
Speaking Task 3	R1 R2 R3 R4	-.03 -.01 .00 .03	.04 .04 .04 .04	.84 1.03 1.01 1.11	.82 1.02 1.01 1.12	M=.00 , SD=.02 , Separation= .00, fixed Chi-Square (df)=1.0(3) , p=.79.
Speaking Task-1	R1 R2	-.01 .01	.04 .04	.96 1.03	.98 1.02	M= .00, SD=.01, Separation=.00, fixed Chi-Square (df) = .1 (1), p=.81.
Speaking Task-2	R1 R2	.00 .00	.03 .03	1.06 .94	1.06 .94	M= .00, SD=.01, Separation=.00, fixed Chi-Square (df) = 5.42 (1), p=.95.
Speaking Task-3	R1 R2	.00 .00	.03 .03	.91 1.06	.93 1.08	M= .00, SD=.00, Separation=.00, fixed Chi-Square (df) = 6.28 (1), p=.91.

Note: R1 to R4= Labels for Rater 1 to Rater 4, MnSq = mean square, M= Mean of Rater Logits per group, SD=standard deviation of rater logit per group, df= degree of freedom, p (value) =significance level.

With regard to writing rater severity, the variable maps shows that rater severity measures (levels) in each of the six speaking dataset are all located on the horizontal level with 0 logit value. This indicates that the severity level across the four (or two) raters within each speaking group were trivial. According to statistics in the third column in Table 3, the severity spans between the most lenient rater and the most severe rater ranges from .00 logit in the last two speaking groups to .27 logits in the Speaking Task 1 group. This relatively large span was produced by Rater 1 (logit=-.15) and Rater 4 (logit=.12). This was

reflected in the group statistics. Of the six separately calibrated groups, the Speaking Task 1 group had the largest separation index value (1.82) as well as the only group having non-zero separation index value. The fixed chi-square (with degree of freedoms) was 17.2 (3), significant at the p=.00 level with rater severity in all other groups displaying a non significant variation. These results seem to suggest that, despite the relative large gap between Rater 1 and Rater 2 in Speaking Task 1 group, rater severity variation did not exist among the raters when they were rating each group of the other five speaking datasets. The fourth column in Table 3 shows that the level of error was small.

Individual rater severity appropriateness was examined by referring to rater infit and rater outfit in the fifth column in Table 3. According to the fifth column, individual rater infit values ranged from .82 (by Rater 1 in Speaking Task 2) to 1.13 (by Rater 4 in peaking Task 2) and the outfit values range from .82 to 1.14 by the same pair of raters. Either through applying the wide or narrow range of lower and upper limits control, all rater severity levels are acceptable. Rater 1 in Speaking Task 1, the previously identified relatively lenient rater, had an infit and outfit of .97 and .94 respectively. Rater 4 in Speaking Task 1, the previously identified relatively harsh rater, had an infit and outfit of 1.05 and 1.08, respectively. These suggest that despite the span, their rating severity or leniency levels were within acceptable. Now it can be concluded that all these raters were internally consistent when rating IELCA speaking tasks.

12. Psychometric Dimension of Criteria

Question 2 in section 3 is addressed here which is divided into two sub questions:

Question 2.1: What is the psychometric dimension of IELCA writing tasks? Are the criteria clearly distinguishable?

Table 4 presents the results of the FACETS analysis for writing criteria calibration: a) task achievement, b) coherence and cohesion, c) lexical resources and d) grammatical range and accuracy. It shows the criterion types, criteria difficulty measures, error and infit and outfit mean-square values. The psychometric quality of the eight IELCA writing datasets were assessed by examining the data-model fit indexes. Using the same fit approach, an infit mean square value of 1.0 indicates perfect fit between the actual ratings on average and the expected ratings by the model. A value less than .70 suggests overfit (or over-predictable from each other) and a value larger than 1.3 implies there is a noticeable noisy component in the ratings (Linacre, 1998)— either case degrades the prevision of the measures (Myford & Engelhard Jr, 2001).

Table 4
Criteria Measurement Report for IELCA Writing Subtest

	Criteria	Criteria Difficulty Measure (in logits)	Error	Fit		Separation Statistics*
				Infit MnSq	Outfit MnSq	
W-AC Task 1	GRA	-.53	.04	1.05	1.00	M= .00, SD= .66, Separation =12.66, fixed Chi-Square (df)= 513.2(3), *p*=.00.
	LR	-.44	.04	.94	.95	
	COCO	.05	.04	1.06	1.04	
	TA	.91	.06	.96	.93	
W-AC Task 2	TA	-.36	.05	1.27	1.25	M=.00, SD=.21, Separation=6.07, fixed Chi-Square (df)= 105.9(3), p=.00.
	LR	-.05	.03	.89	.89	
	GRA	.14	.03	.90	.90	
	COCO	.17	.03	.96	.96	
W-AC Task-1	GRA	-1.34	.10	.91	.91	M=.00, SD= .1.62, Separation=14.69, fixed Chi-Square (df)=14.69 (3), p= .00.
	LR	-.84	.10	.84	.87	
	COCO	-.58	.10	.93	.95	
	TA	2.76	.14	1.25	1.14	
W-AC Task-2	GRA	-.32	..05	.89	.88	M=.00 , SD=.42 , Separation= 6.87, fixed Chi-Square (df)=137.9(3) , p=.00.
	LR	-.22	.05	.87	.79	
	COCO	-.18	.05	.84	.77	
	TA	.72	.08	1.45	1.57	
W-GT Task 1	TA	-.34	.05	1.16	1.18	M=.00 , SD=.21 , Separation=5 .00, fixed Chi-Square (df)=79.0 (3) , p=.00
	GRA	.02	.04	1.02	.97	
	LR	.11	.04	.85	.82	
	COCO	.21	.04	.98	.96	
W-GT Task 2	LR	-.30	.03	.71	.69	M= .00, SD=.42 , Separation=12.56 ,fixed Chi-Square (df)=442.3(3) , p=.00
	COCO	-.22	.03	.85	.87	
	GRA	.20	.03	1.16	1.08	
	TA	.73	.04	1.25	1.27	
W-GT Task-1	COCO	-.31	.07	.94	.89	M= .00, SD=.39, Separation=5.13, fixed Chi-Square (df) =76.8 (3), p=.00.
	GRA	-.23	.06	1.15	1.05	
	LR	-.12	.06	.82	.78	
	TA	.66	.10	1.09	1.05	
W-GT Task-2	TA	-.25	.08	1.20	1.22	M= .00, SD=.19, Separation=3.30 , fixed Chi-Square (df)=45.19(3) , p=.00.
	LR	-.07	.05	.87	.83	
	COCO	.05	.05	1.08	1.03	
	GRA	.28	.05	.82	.79	

Note: W-AC= writing for academic purpose, W-GT=writing for general training purpose , R1 to R4= Labels for Rater 1 to Rater 4, MnSq = mean square, M= Mean of Rater Logits per group, SD=standard deviation of rater logit per group, df= degree of freedom, p (value) =significance level, GRA=grammatical range and accuracy, TA=task achievement, LR=lexical resources, COCO=cohesion

and coherence.

Referring to the left side of the fifth column in Table 4, the infit values for task achievement ranged from .96 to 1.27; those for coherence and cohesion ranged from .84 to 1.08; those for lexical resources ranged from .71 to .94; and those for grammatical range and accuracy ranged from .82 to 1.16. As all these values are within the narrow quality limit control of .70 and 1.30, there was no evidence of multidimensionality in each of the eight IELCA writing tasks. In addition, as presented in the last column of Table 4, the separation indices were all larger than 5.00 and all fixed chi-square indices were significant at p=.00 level. These results imply that, on the one hand, the null hypothesis on ratings of the four scoring criteria did not involve other dimensions which could not be rejected; and on the other, despite all four criteria functioning consistently during the rating process, the four criteria were far from being homogeneous.

Question 2.1: What is the psychometric dimension of IELCA speaking tasks? Are the criteria clearly distinguishable?

Table 5 presents the results of the FACETS analysis for speaking criteria calibration: a) pronunciation, b) fluency, c) language accuracy and appropriacy, and d) task fulfillment. It shows the criterion types, criteria difficulty measures, error and infit and outfit mean-square values. As with writing, the psychometric quality of the six IELCA speaking datasets were assessed by examining the data-model fit indexes.

As shown in the left side of the fifth column in Table 5, the infit values for pronunciation ranged from 1.00 to 1.22; those for fluency ranged from .85 to 1.09; those for language accuracy and appropriacy ranged from .86 to 98; and those for task fulfillment ranged from .84 to 1.13. As all these values are within the narrow quality limit control of .70 and 1.30, there was no evidence of multidimensionality in each of the eight IELCA writing tasks. Table 5 also presents separation indices regarding the four criteria. The separation indices values were all larger than 15 and all fixed chi-square indices with degrees of freedom accounted were significant at p=.00 level. In all, these results imply the unidimensionality of the four criteria used to rate IELCA speaking and the clear distinction between these criteria.

Table 5
Criteria Measurement Report for IELCA Speaking Subtest

	Criteria	Criteria Difficulty Measure (in logits)	Error	Fit Infit MnSq	Outfit MnSq	Group Statistics*
Speaking Task 1	TA	-.45	.05	1.04	1.05	M= .00, SD= .32,
	FLU	-.01	.05	.90	.88	Separation =6.37,

	AA	.00	.05	.93	.95	fixed Chi-Square
	PRO	.46	.05	1.12	1.09	(df)= 165.6(3), p=.00.
Speaking Task 2	TA	-.82	.05	.97	.95	M=.00, SD=.78,
	AA	-.29	.05	.91	.91	Separation=15.84,
	FLU	-.18	.05	1.08	1.06	fixed Chi-Square
	PRO	1.29	.05	1.00	1.04	(df)= 993.8(3), p=.00.
Speaking Task 3	TA	-.64	.04	1.13	1.12	M=.00 , SD=.87 ,
	AA	-.53	.04	.93	.90	Separation=19.82,
	FLU	.33	.04	.85	.85	fixed Chi-Square
	PRO	1.50	.05	1.09	1.09	(df) = 1493.4 (3) , p=.00.
Speaking Task-1	TA	-.60	.05	.84	.84	M= .00, SD=.50,
	FLU	-.18	.05	1.09	1.11	Separation=9.88,
	AA	-.01	.05	.86	.90	fixed Chi-Square
	PRO	.78	.05	1.18	1.14	(df) = 393.5 (3), p=.00.
Speaking Task-2	TA	-.73	.05	.88	.88	M= .00, SD=.73,
	AA	-.32	.05	.98	.99	Separation=15.08,
	FLU	-.17	.05	.91	.91	fixed Chi-Square
	PRO	1.22	.05	1.22	1.22	(df) =904.9(3), p=.00.
Speaking Task-3	TA	-.64	.04	.93	.94	M= .00, SD=.73,
	AA	-.39	.04	.89	.91	Separation=.16.09,
	FLU	-.20	.04	.99	.99	fixed Chi-Square
	PRO	1.23	.05	1.16	1.18	(df) = 990.7(3), p=.00.

Note: R1 to R4= Labels for Rater 1 to Rater 4, MnSq = mean square, M= Mean of Rater Logits per group, SD=standard deviation of rater logit per group, df= degree of freedom, p (value) =significance level, TA=task achievement, accuracy and appropriacy, FLU=fluency, PRO=pronunciation.

13. Discussion and conclusion

In this report, the researcher used the many-facet Rasch measurement and explored two sources of variability (rater and criteria domains) in IELCA writing and speaking scores. The global fit results showed that all IELCA writing and speaking tasks sufficiently fit the three-facet Rasch model (i.e., examinee, rater and criteria).

The investigation into rater facet showed that rater severity did not exist among raters when they were rating IELCA writing and speaking performances. This seems to confirm findings in L2 performance rating research that rater training

can help increase inter-rater and intra-rater reliability (McNamara, 1996).

The probe into criteria facet revealed that each of the four scoring criteria (writing and speaking) work in concert among themselves. The correspondence of ratings on one criterion to ratings on another indicates a single pattern of writing (or speaking) across all criteria on the same rubric. This feature makes the combination of scores on different criterion meaningful. Furthermore, despite the writing (or speaking) criteria domains function being consistent in a single pattern, they were also clearly distinguished from each other rather than redundant of the other.

To conclude, this study used the many-facet Rasch measurement to explore potential variance from two sources: rater and scoring criteria. The results did not identify variance from rater or scoring criteria facets in IELCA writing or speaking subtests. These many-facet Rasch analysis results, though merely based test scores, did serve as evidence for the construct validity of IELCA writing and speaking subtests.

14. References

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Crocker, L. (1997). Assessing content representativeness of performance assessment exercises. *Applied Measurement in Education, 10*(1), 83-95.

Eckes, Thomas. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly: An International Journal, 2*(3), 197-221.

Linacre, J.M. (1998). Investigating rating scale category utility. *Journal of Outcome Measurement, 3*(2), 103-122.

Linacre, J.M. (2002). What do infit and outfit, mean-square and standardized mean? . *Rasch Measurement Transactions, 16*(2), 878.

Linacre, J.M. (2013a). Facets computer program for many-facet Rasch measurement, version 3.71. 0. Beaverton, Oregon: Winsteps. com.

Linacre, J.M. (2013b). A user's guide to FACETS Rasch-model computer programs.

Linacre, J.M., & Wright, B.D. (2002). Construction of measures from many-facet data. *Journal of Applied Measurement, 3*(4), 486-512.

McNamara, T.F. (1996). *Measuring second language performance*. New York: Longman.

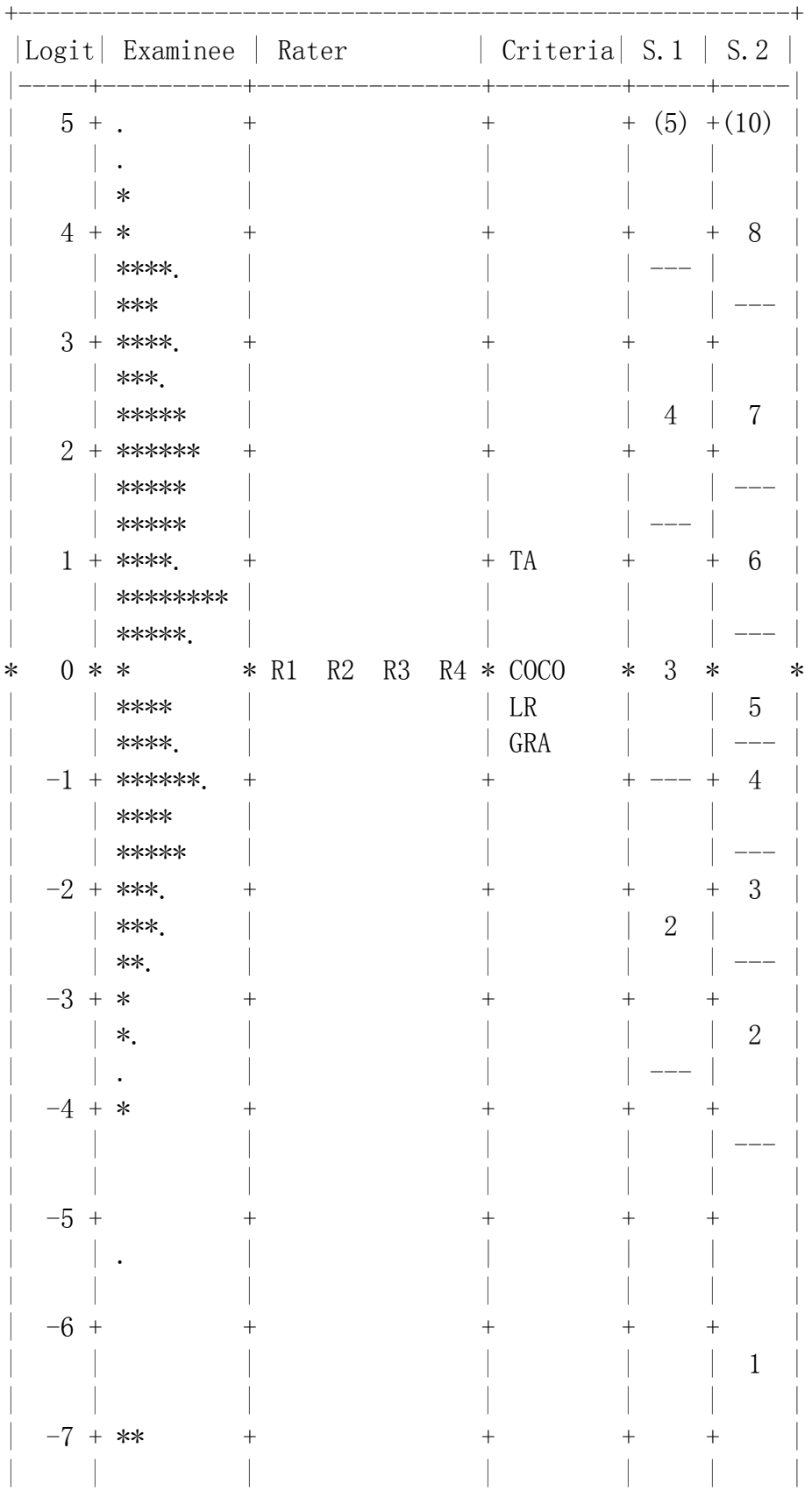
Myford, C.M., & Engelhard Jr, George. (2001). Examining the Psychometric Quality of the National Board for Professional Teaching Standards Early Childhood/Generalist Assessment System*. *Journal of Personnel Evaluation in Education, 15*(4), 253-285.

Myford, C.M., & Wolfe, E.W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*, 386-422.

Weigle, Sara Cushing. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263-287.

Appendix A

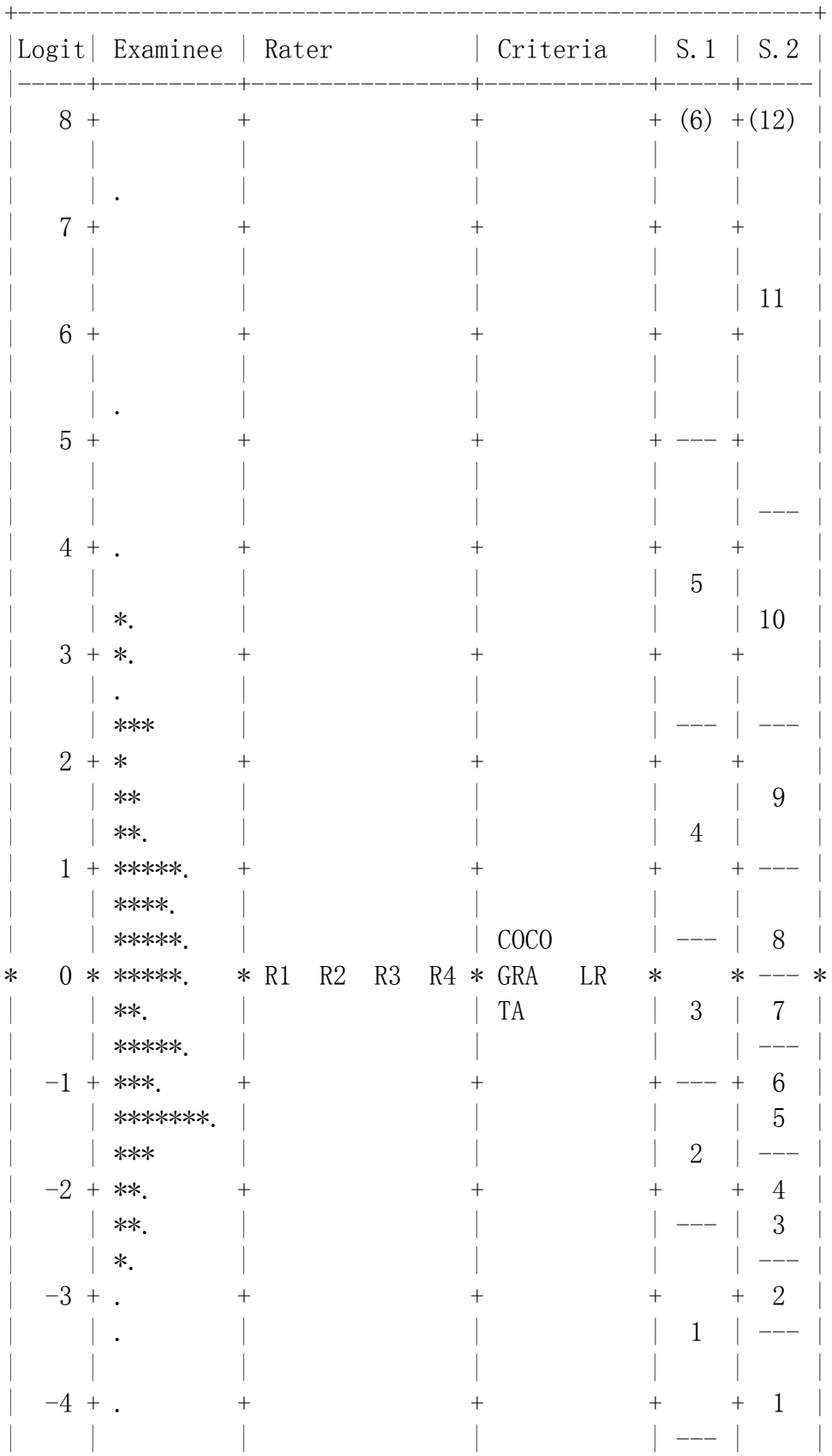
FACETS variable map of IELCA AC writing Task 1 data

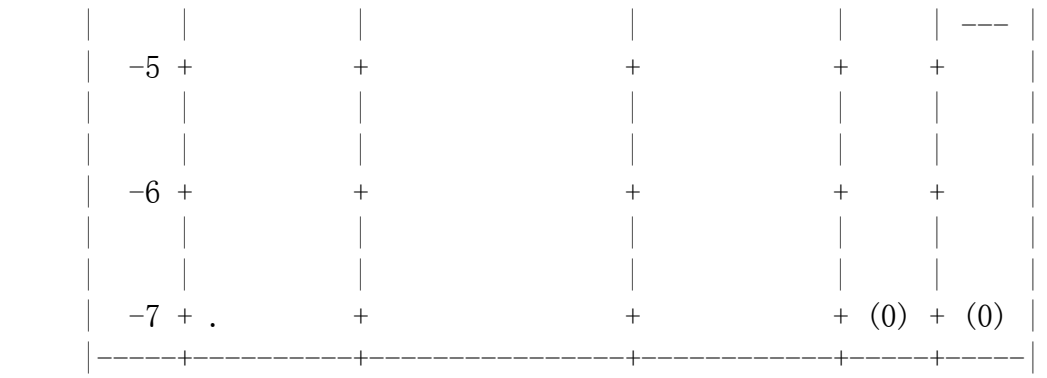


Note: R1 to R4 represents four raters respectively. TA=task achievement, COCO= Cohesion and coherence, LR=lexical resources, GRA=grammar range and accuracy.

Appendix B

FACETS variable map of IELCA AC writing Task 2 data

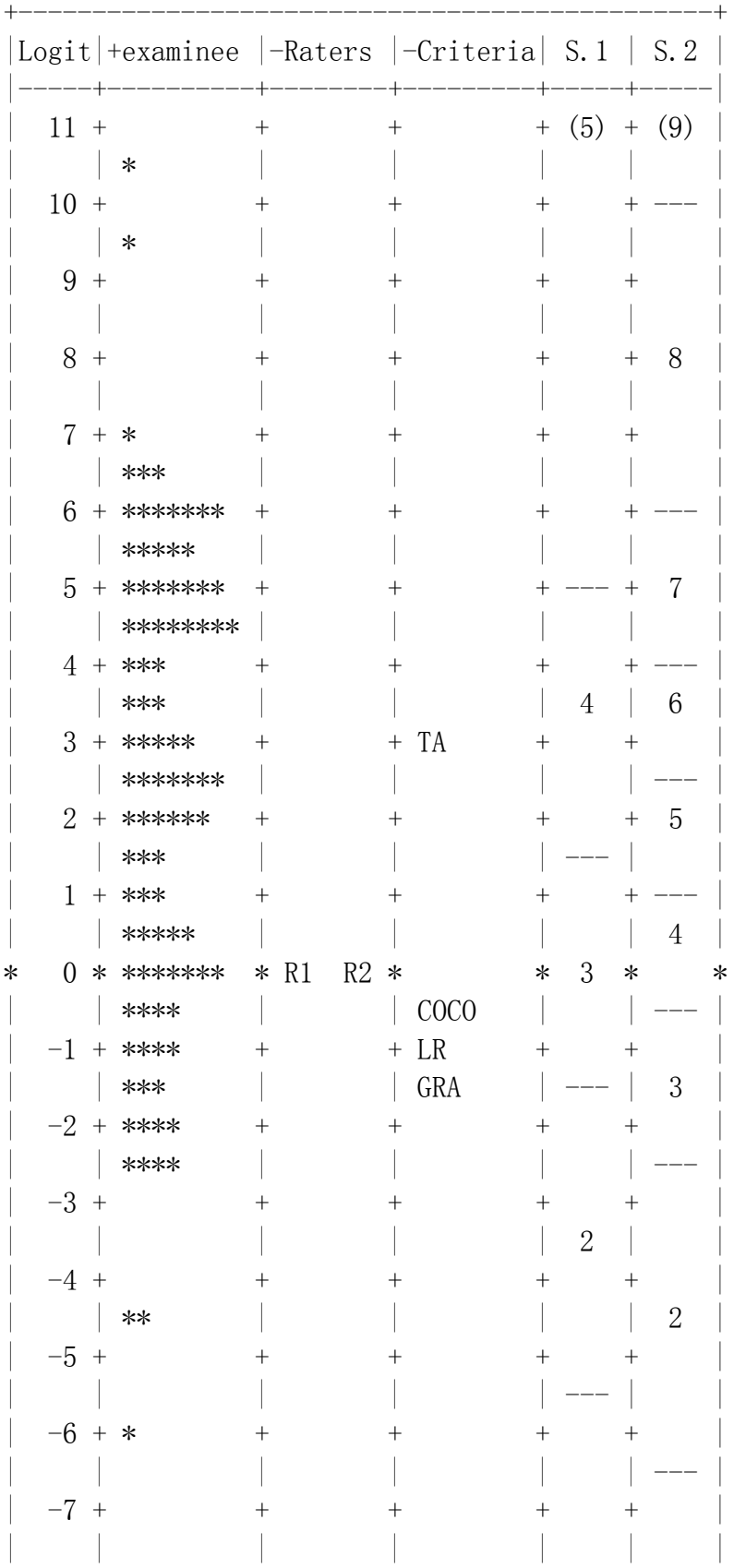


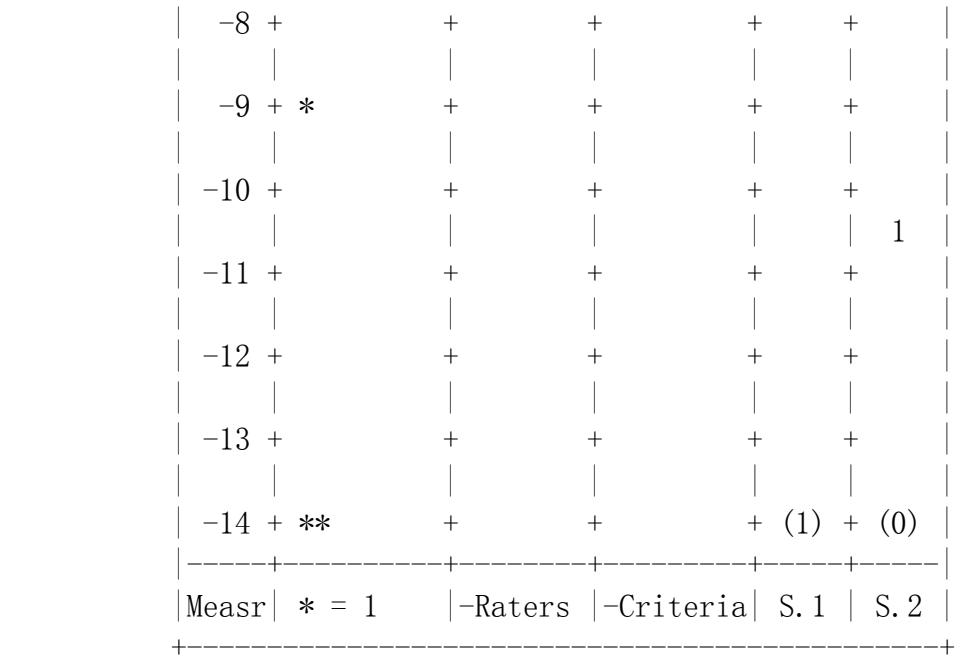


Note: R1 to R4 represents four raters respectively. TA=task achievement, COCO= Cohesion and coherence, LR=lexical resources, GRA=grammar range and accuracy.

Appendix C

FACETS variable map of IELCA AC writing Task -1 data

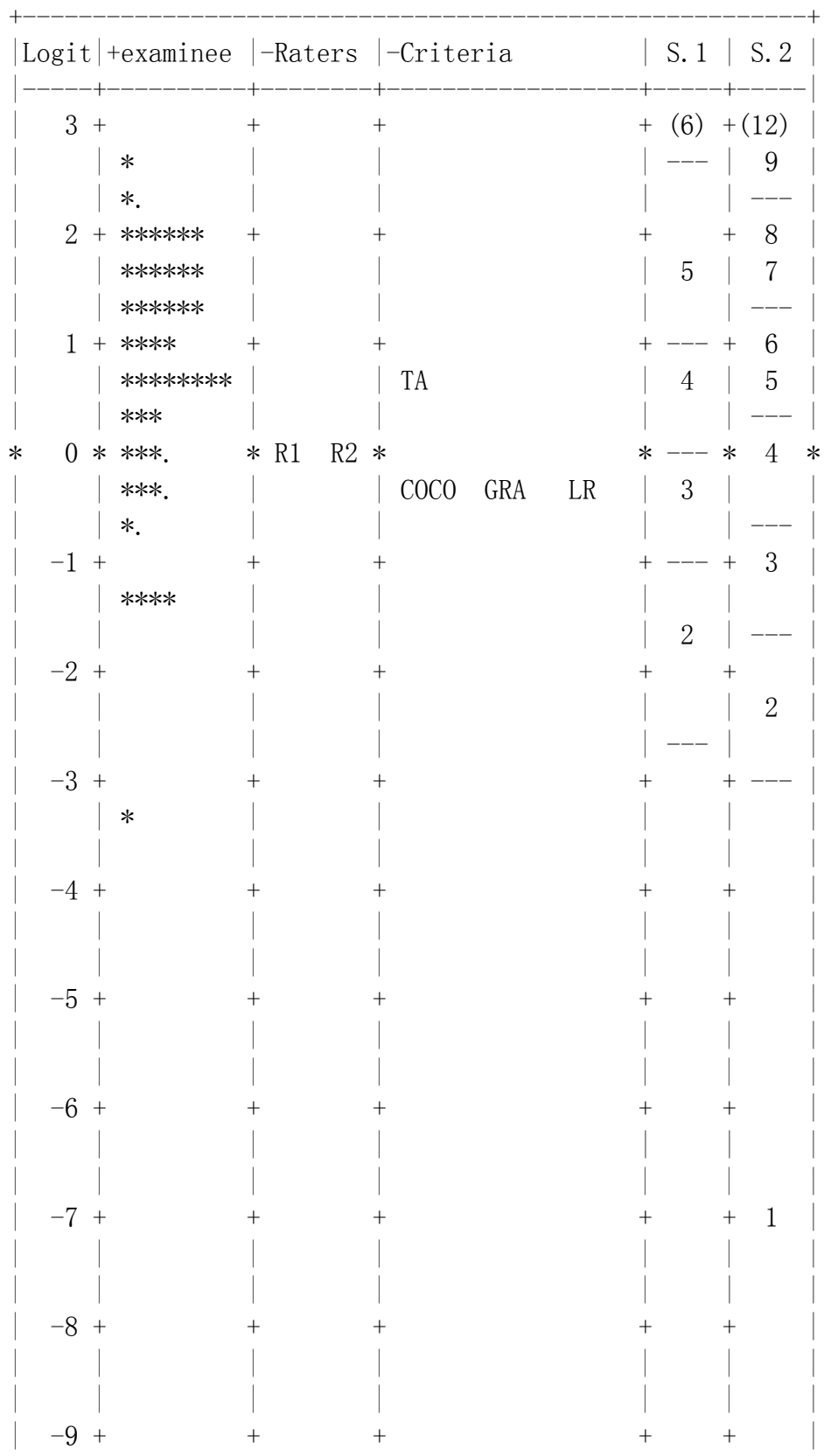


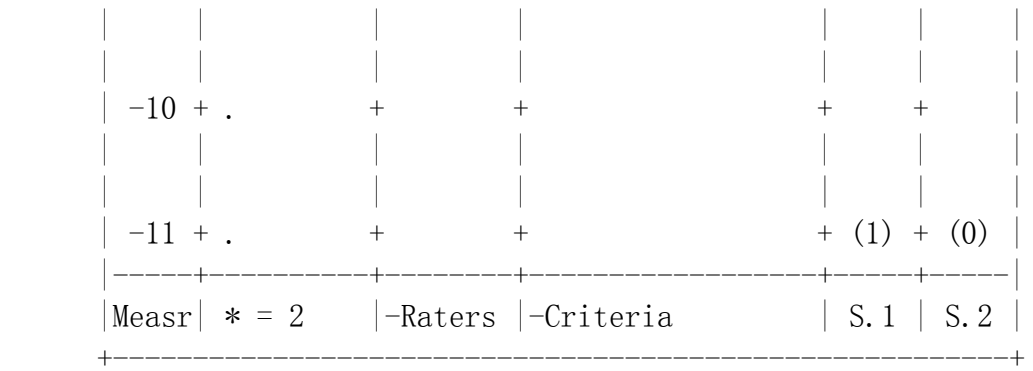


Note: R1=Rater 1, R2=Rater2, TA=task achievement, COCO= Cohesion and coherence, LR=lexical resources, GRA=grammar range and accuracy.

Appendix D

FACETS variable map of IELCA AC writing Task -2 data

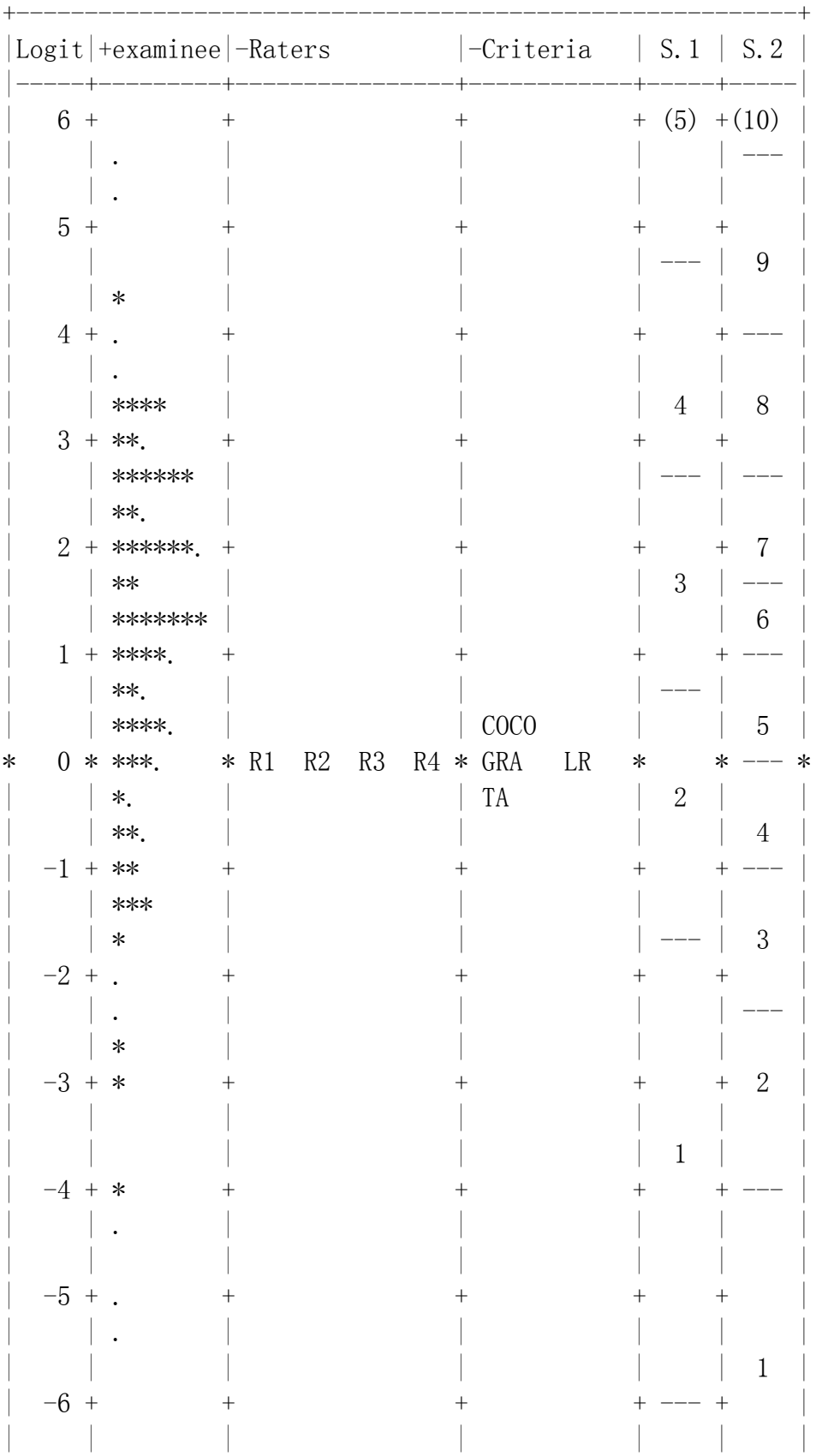


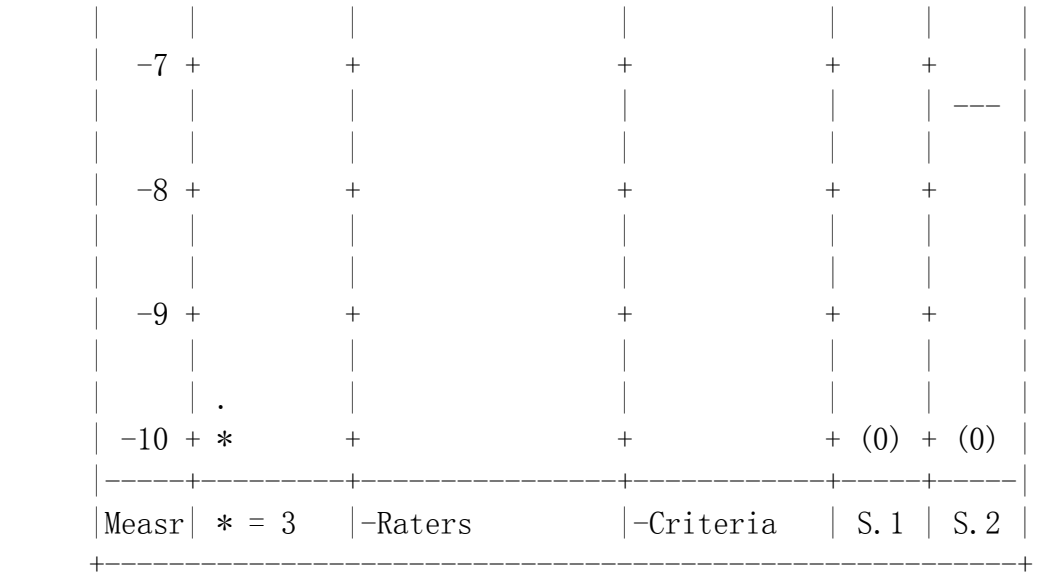


Note: R1=Rater 1, R2=Rater2, TA=task achievement, COCO= Cohesion and coherence, LR=lexical resources, GRA=grammar range and accuracy.

Appendix E

FACETS variable map of IELCA GT writing Task 1 data

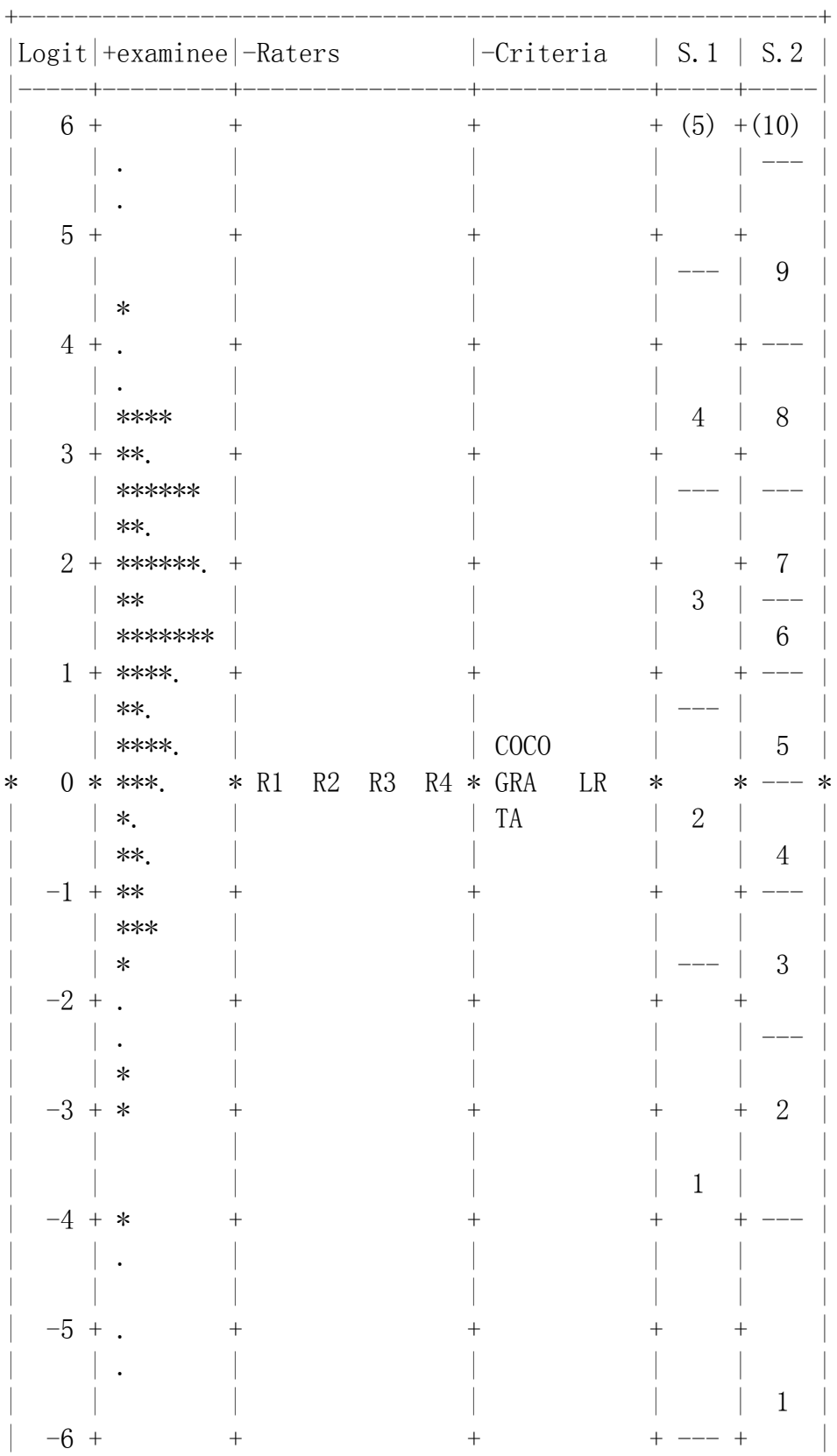


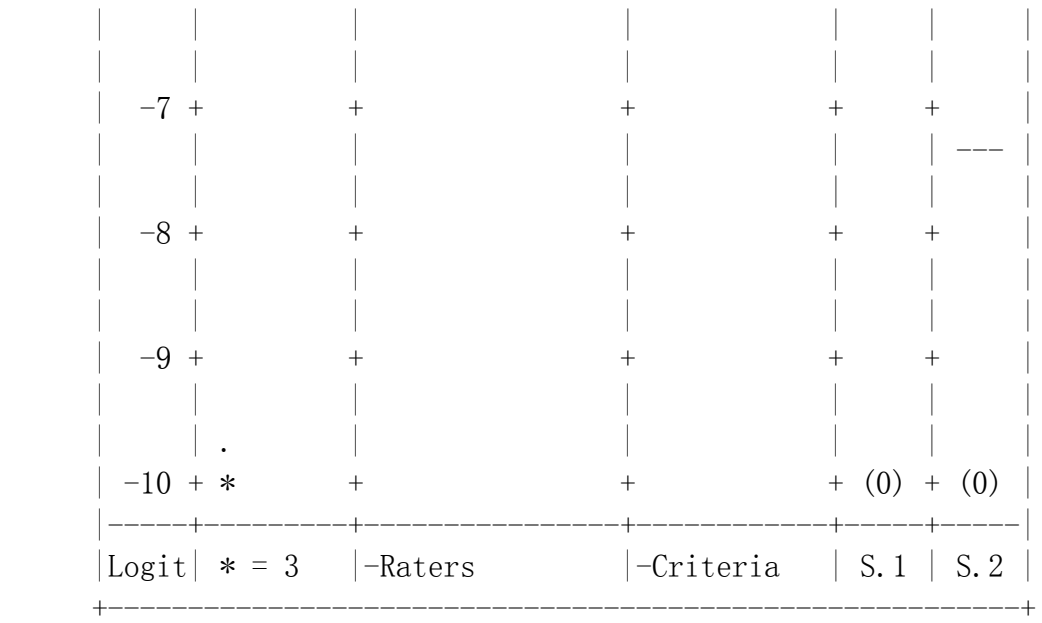


Note: R1 to R4= Rater 1 to Rater 4, TA=task achievement, COCO= Cohesion and coherence, LR=lexical resources, GRA=grammar range and accuracy.

Appendix F

FACETS variable map of IELCA GT writing Task 2 data

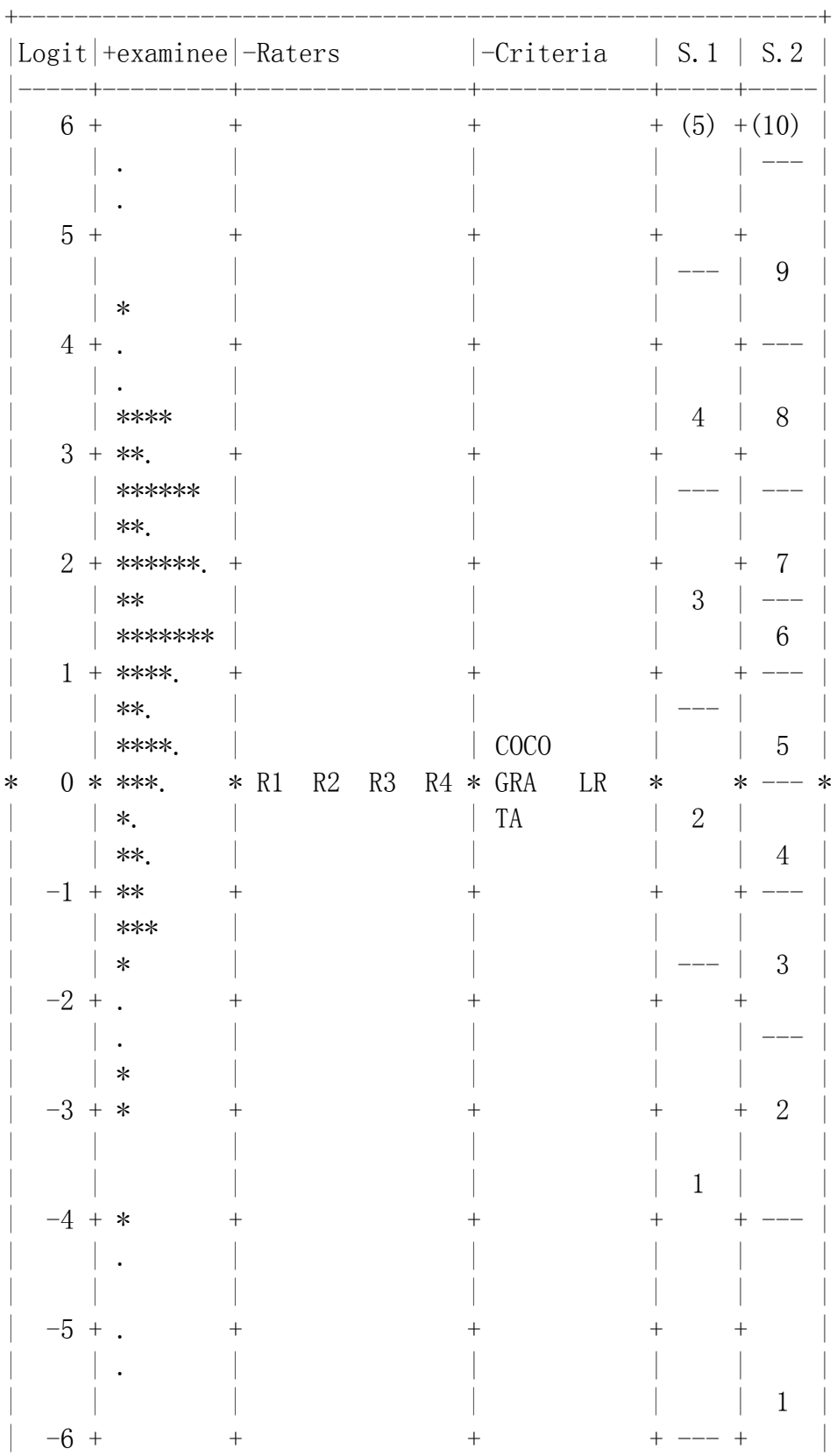


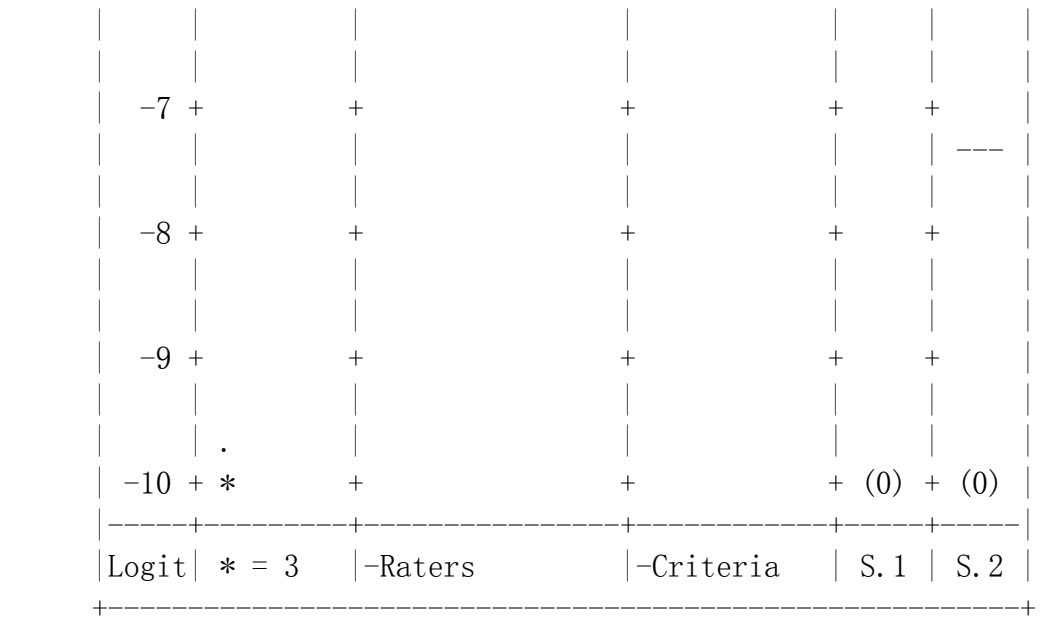


Note: R1 to R4= Rater 1 to Rater 4, TA=task achievement, COCO= Cohesion and coherence, LR=lexical resources, GRA=grammar range and accuracy.

Appendix G

FACETS variable map of IELCA GT writing 1 data

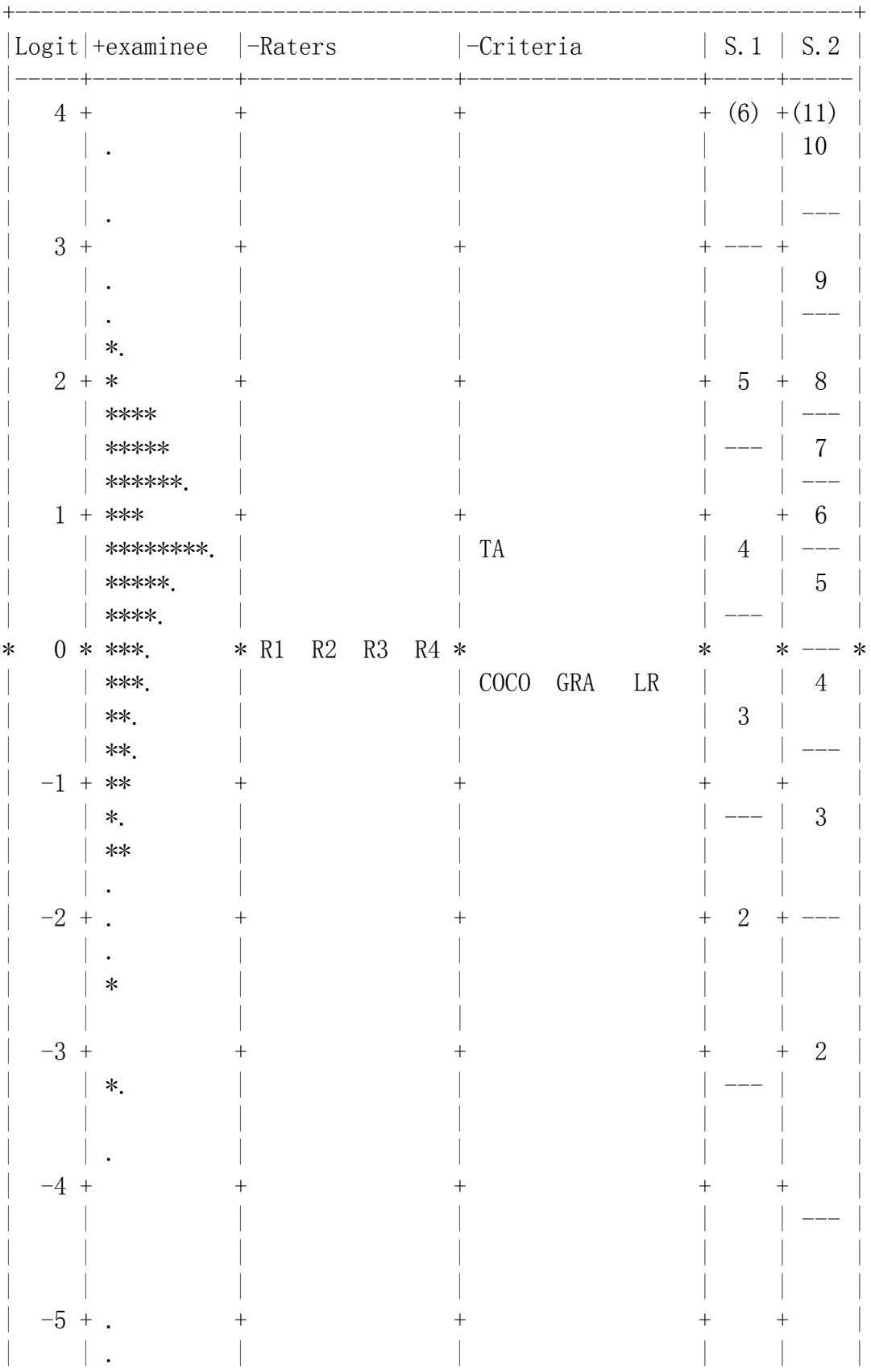


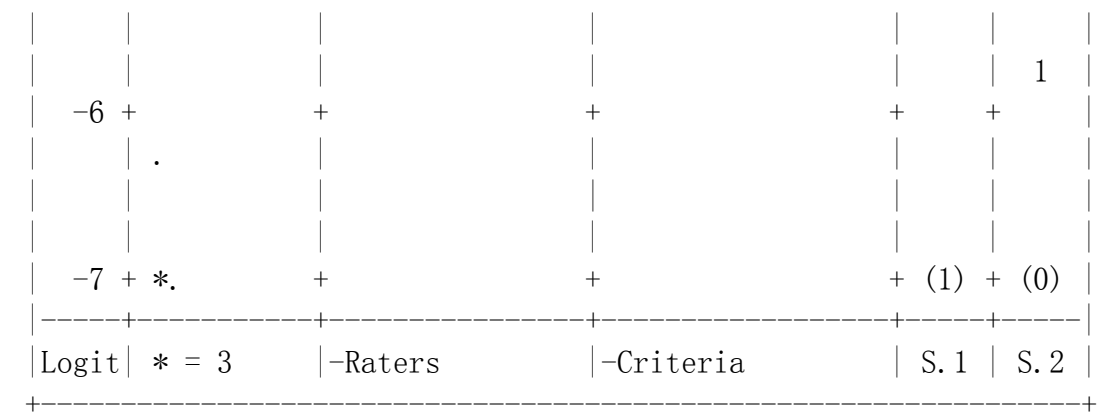


Note: R1 to R4= Rater 1 to Rater 4, TA=task achievement, COCO= Cohesion and coherence, LR=lexical resources, GRA=grammar range and accuracy.

Appendix H

FACETS variable map of IELCA GT writing 2 data

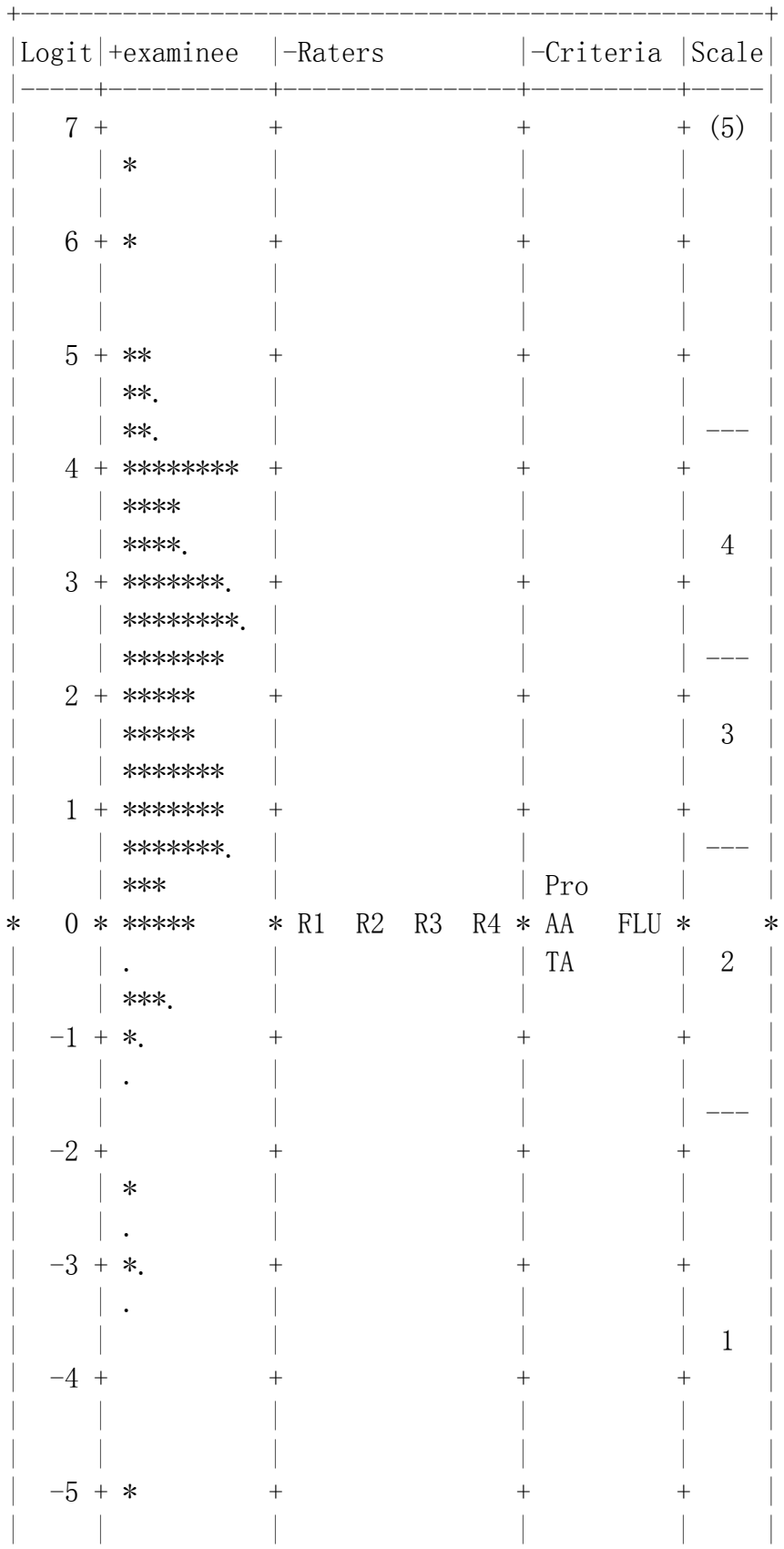


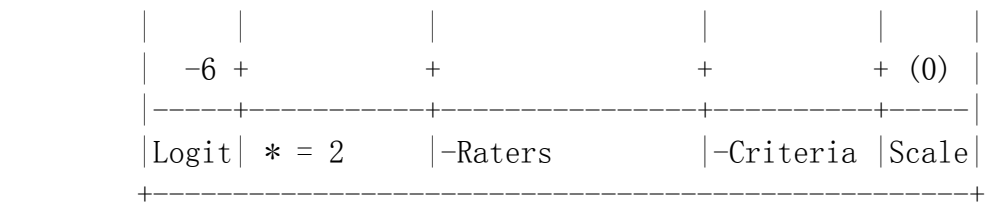


Note: R1 to R4= Rater 1 to Rater 4, TA=task achievement, COCO= Cohesion and coherence, LR=lexical resources, GRA=grammar range and accuracy.

Appendix I

FACETS variable map of IELCA Speaking Task 1 data

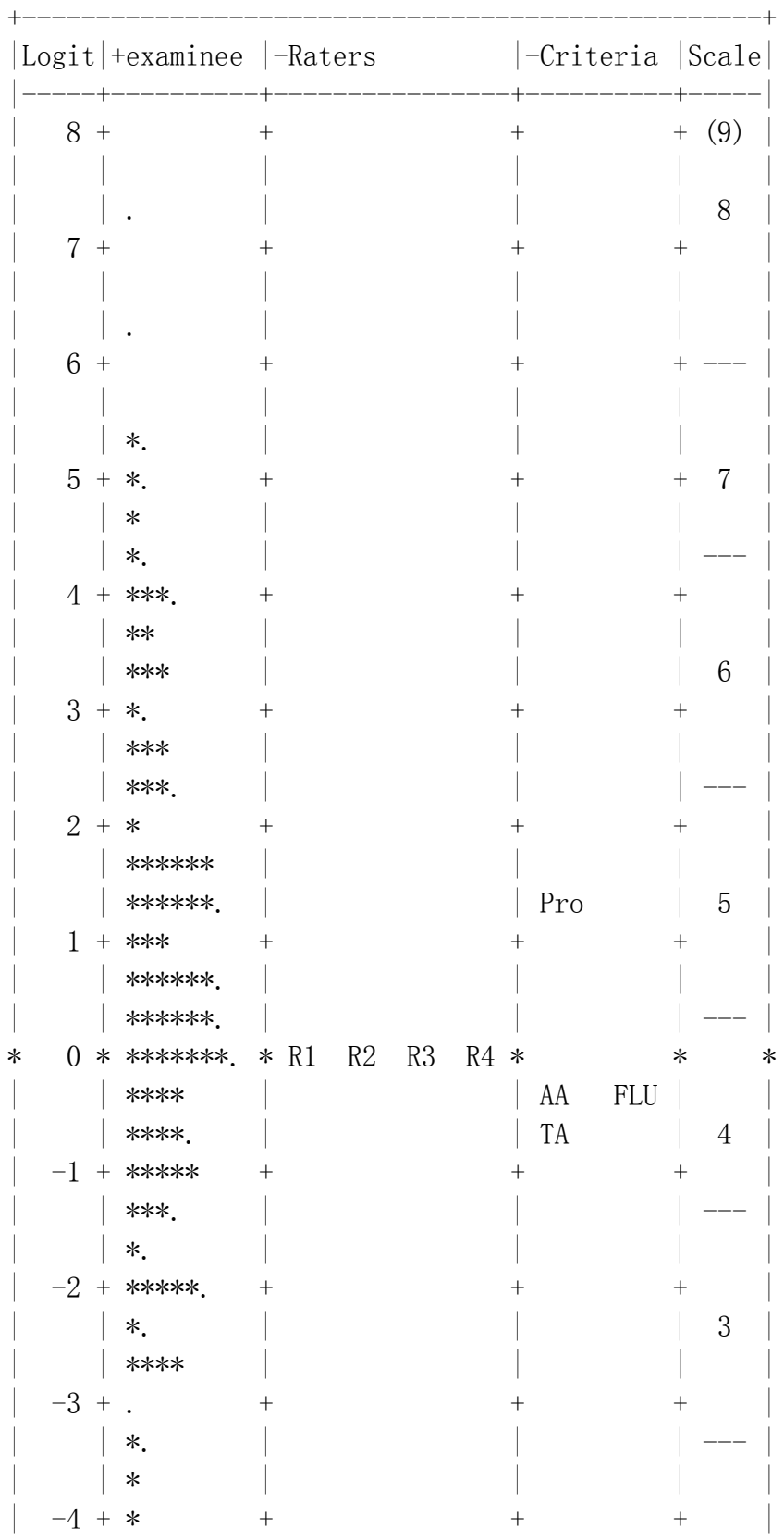


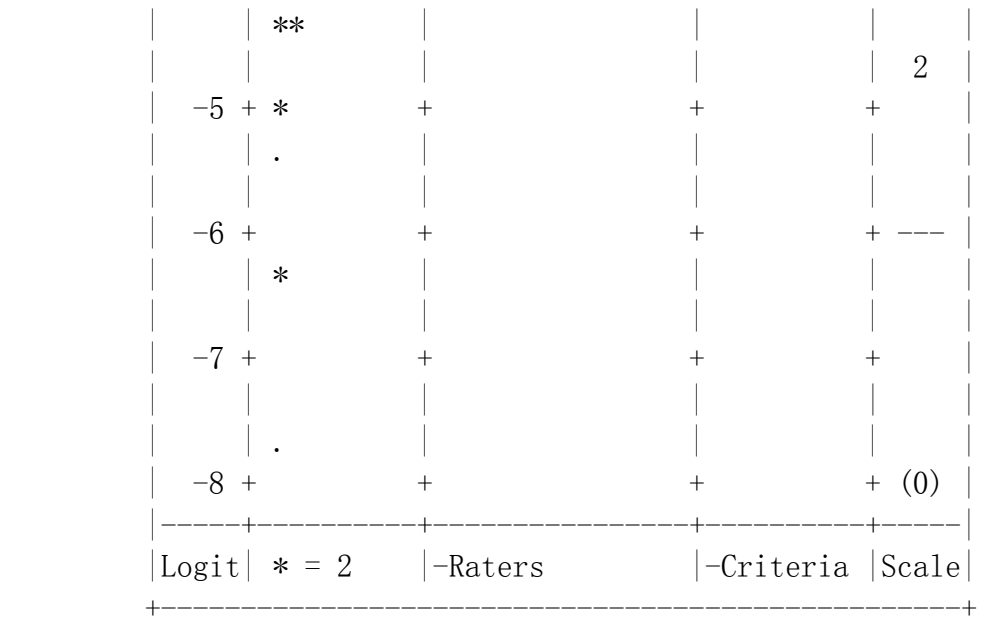


Note: R1 to R4= Rater 1 to Rater 4, PRO=pronunciation, AA=accuracy and appropriacy, FLU=fluency, TA=task achievement

Appendix J

FACETS variable map of IELCA Speaking Task 2 data

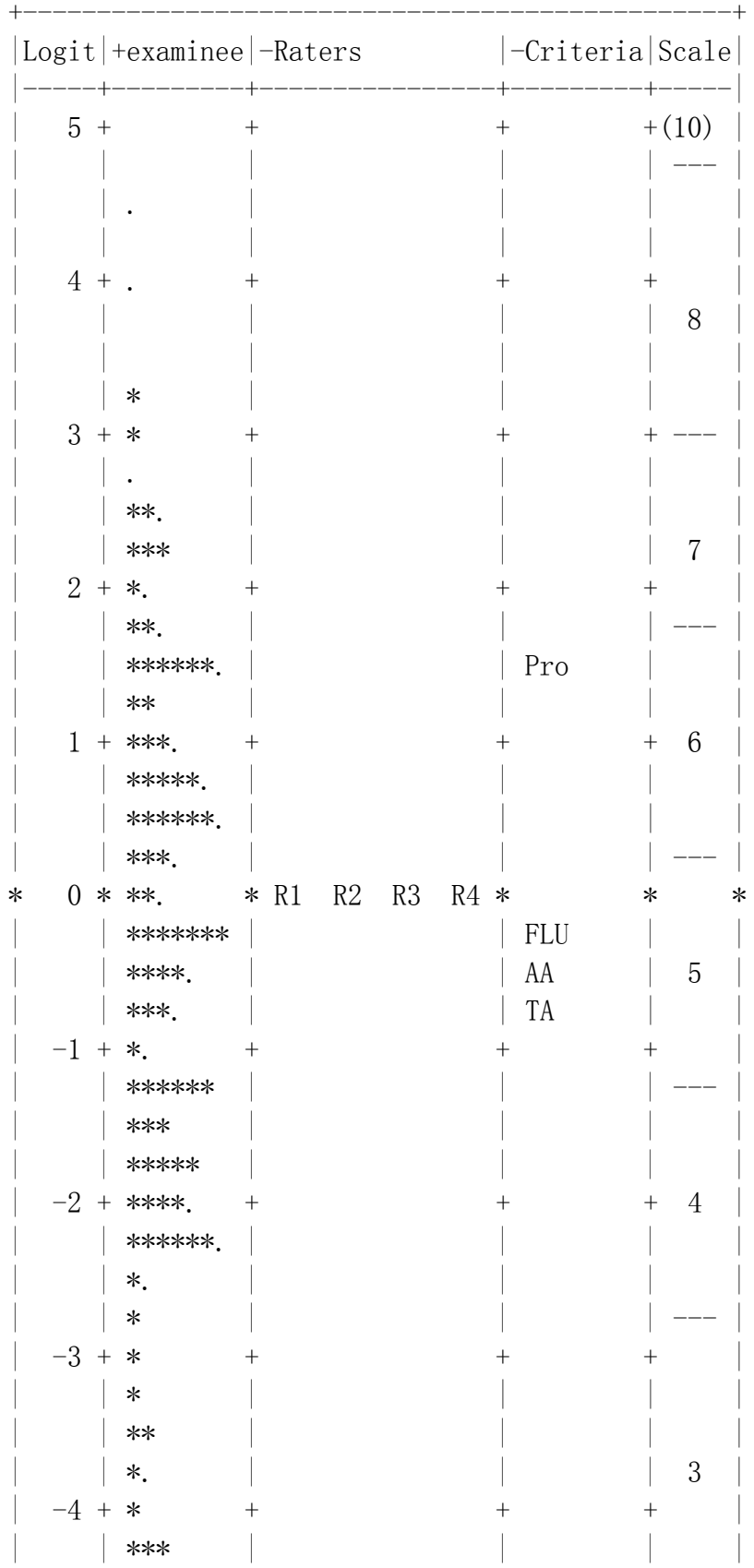


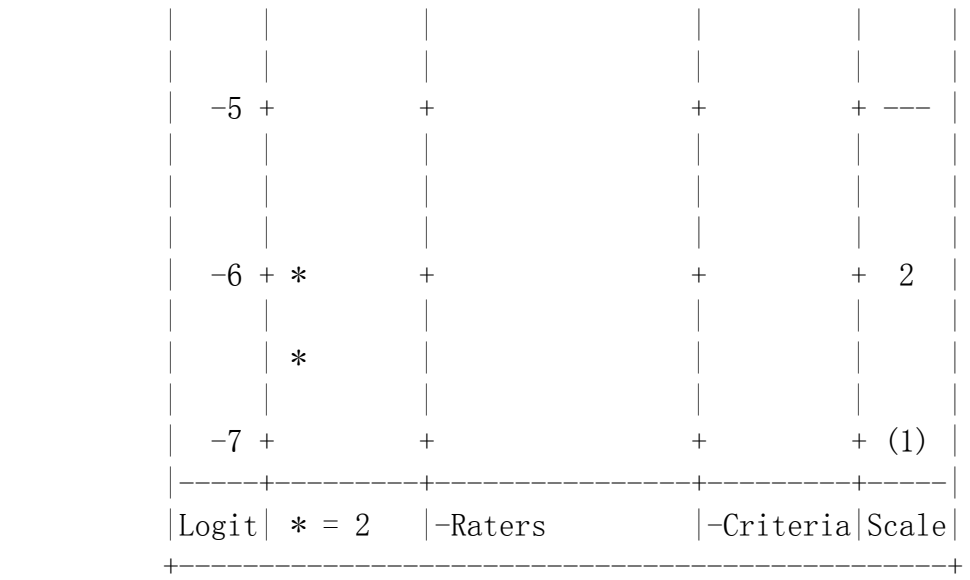


Note: R1 to R4= Rater 1 to Rater 4, PRO=pronunciation, AA=accuracy and appropriacy, FLU=fluency, TA=task achievement

Appendix K

FACETS variable map of IELCA Speaking Task 3 data

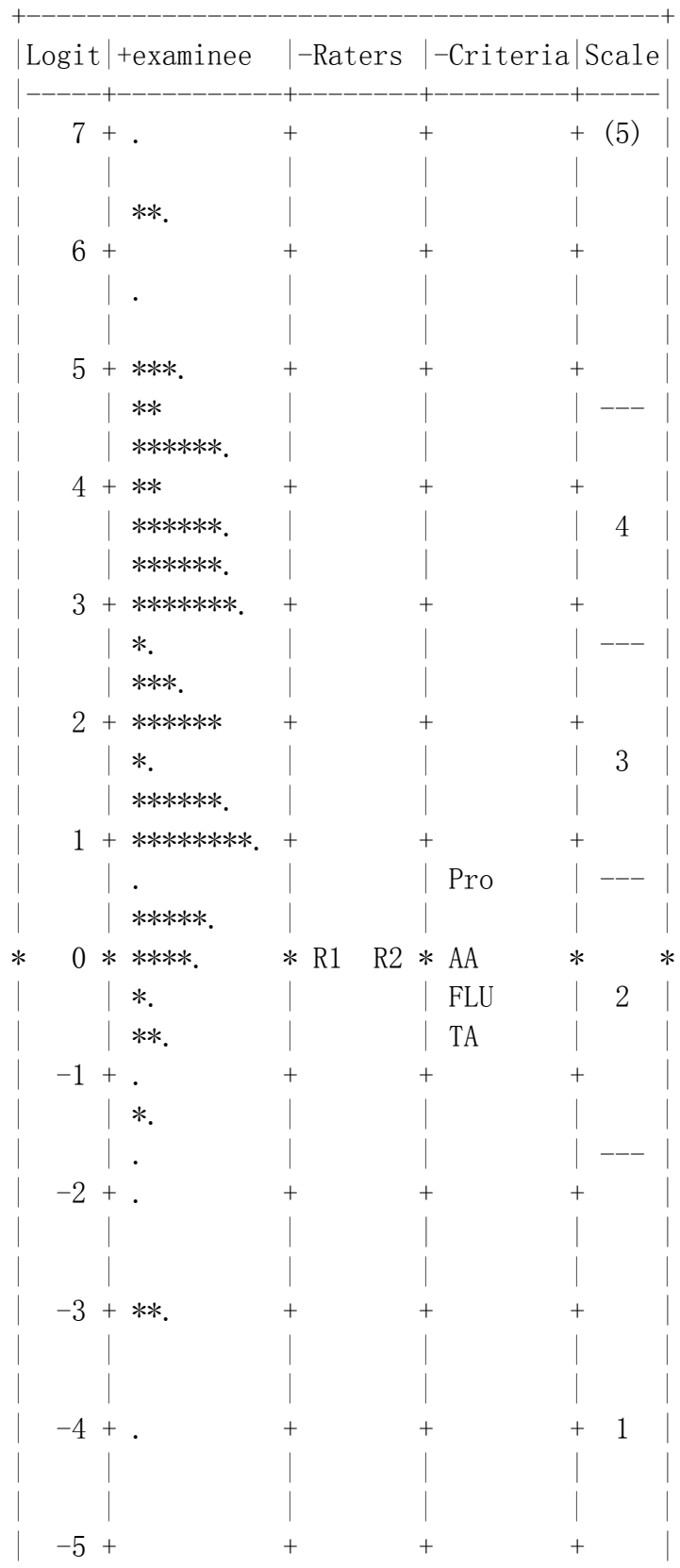


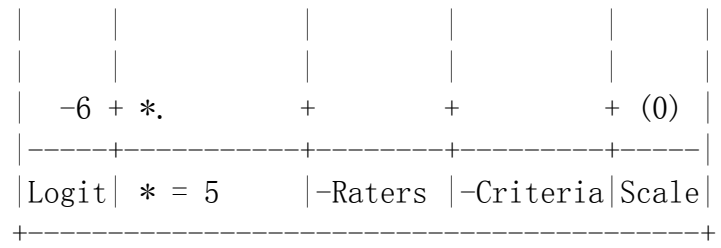


Note: R1 to R4= Rater 1 to Rater 4, PRO=pronunciation, AA=accuracy and appropriacy, FLU=fluency, TA=task achievement

Appendix L

FACETS variable map of IELCA Speaking Task-1 data

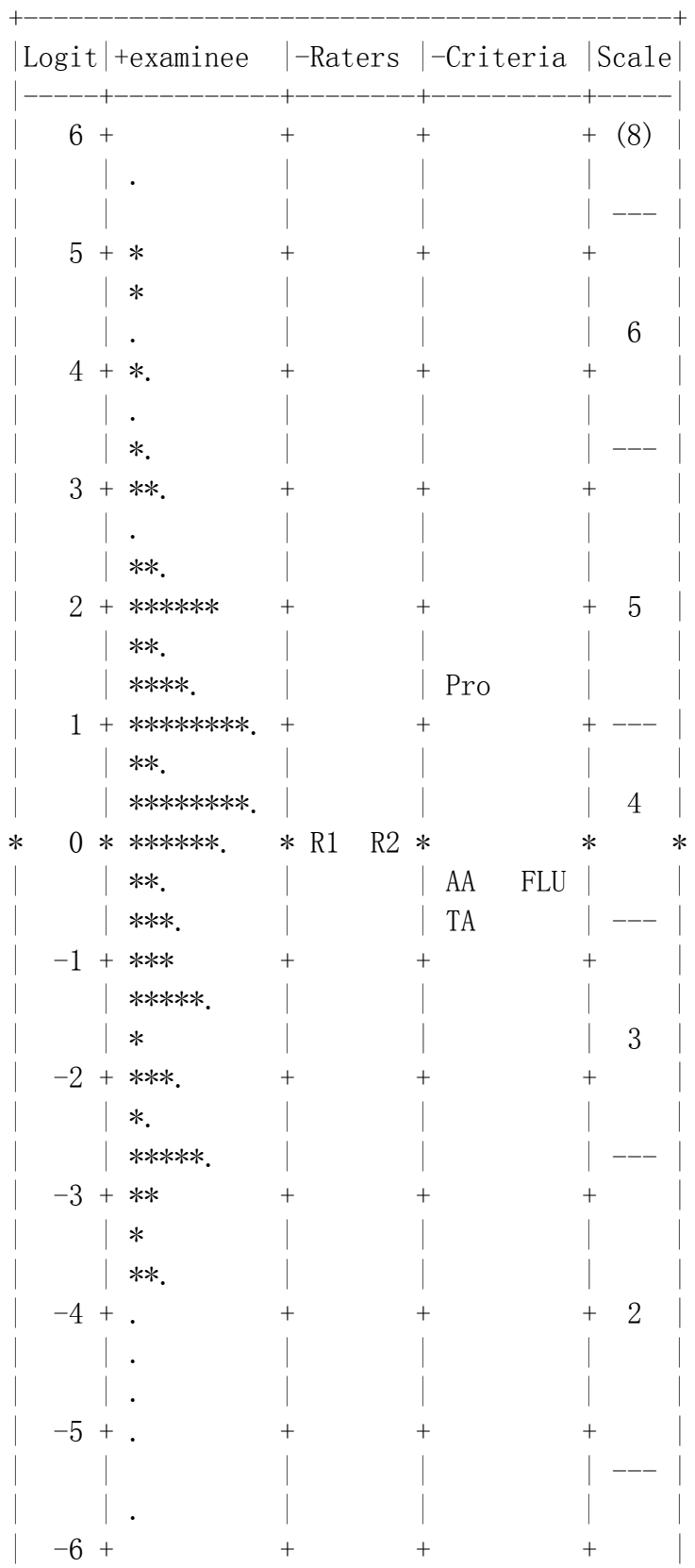


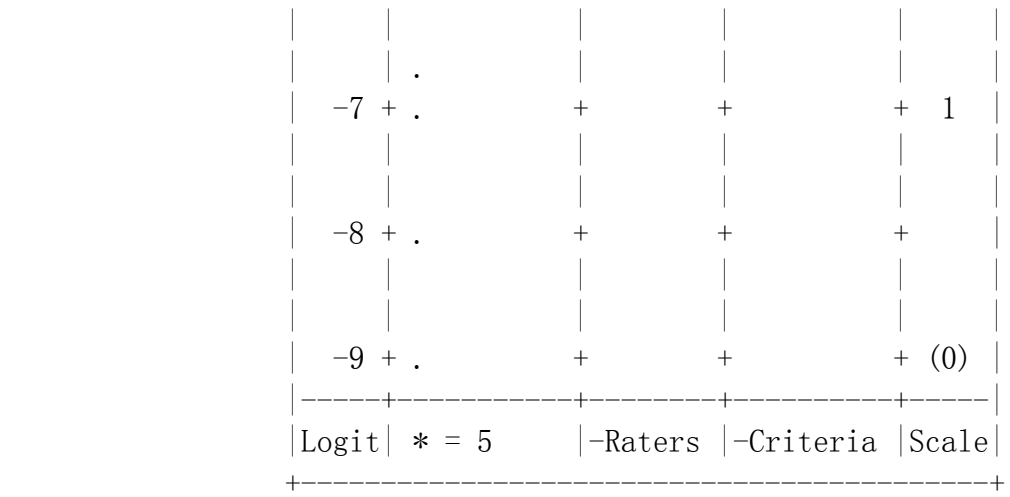


Note: R1= Rater 1, R2=Rater 2, PRO=pronunciation, AA=accuracy and appropriacy, FLU=fluency, TA=task achievement

Appendix M

FACETS variable map of IELCA Speaking Task-2 data

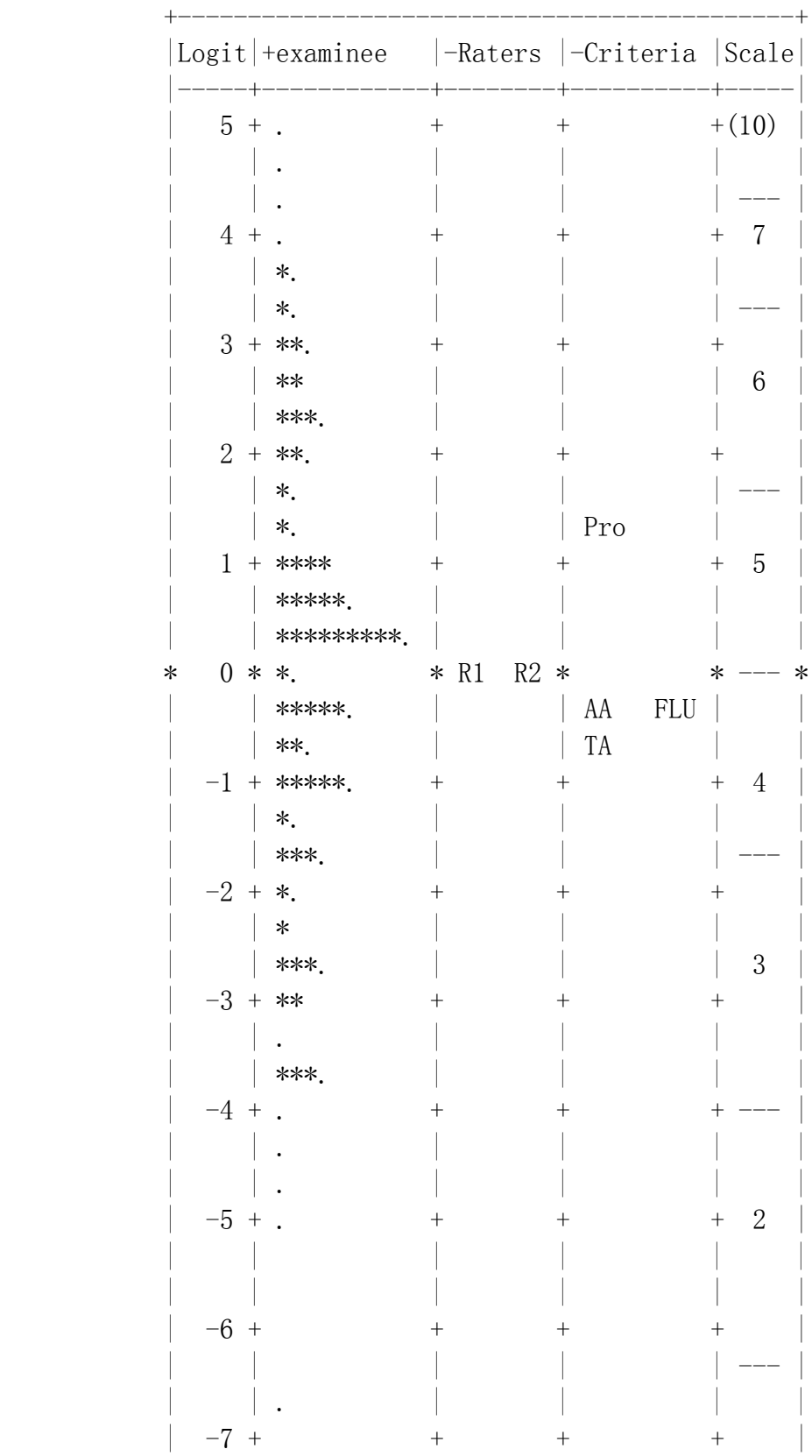


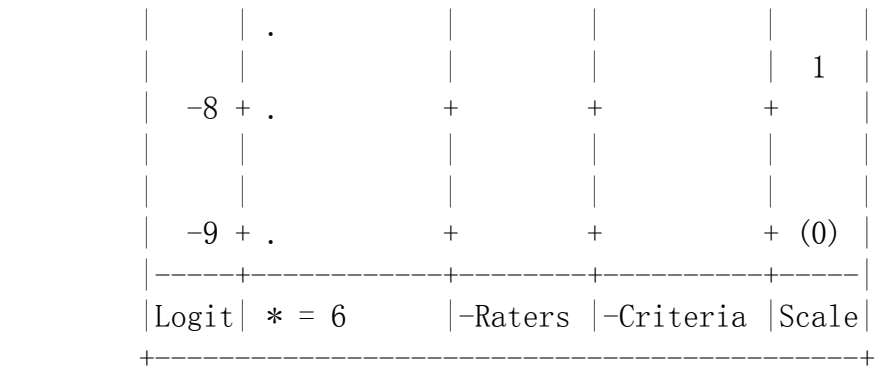


Note: R1= Rater 1, R2=Rater 2, PRO=pronunciation, AA=accuracy and appropriacy, FLU=fluency, TA=task achievement

Appendix N

FACETS variable map of IELCA Speaking Task-3 data





Note: R1= Rater 1, R2=Rater 2, PRO=pronunciation, AA=accuracy and appropriacy, FLU=fluency, TA=task achievement