2017 ELIPT WRITING TEST REVISION PROJECT

JAMES DEAN BROWN, WEI-LI HSU, & KENTON HARSCH

University of Hawai'i at Mānoa

INTRODUCTION

Background

The main purpose of this research project is to revise a previous rating rubric used in the English Language Institute placement test (ELIPT) at the University of Hawaii at Manoa (UHM). UHM requires newly admitted international students, who do not already meet the university's criteria for automatic exemption from the ELI, to take the ELIPT prior to registering for courses in their first semester. Although, when applying for UHM programs, international students report their scores from standardized English proficiency tests, such as TOEFL or IELTS, more fine-grained information is necessary to assess students' language abilities, as well as their needs for support with English for academic purposes during their first few semesters at UHM. In terms of writing proficiency, the ELI has determined that students should be able (a) to fluently generate sufficient written texts, (b) to achieve specific purposes, (c) for identifiable audiences, (d) with effective strategies, (e) at different stages of the writing process (such as generating, revising, editing, and proofreading), and (f) incorporate information from relevant and credible sources according to acceptable citation standards.

The English Language Institute Writing Courses

There are three levels of writing courses provided by the ELI, but the ELIPT-Writing examinees are placed to seven levels, including four levels for undergraduate students and three levels for graduate students. The four undergraduate levels are (a) ELI 73 for intermediate writing, (b) ESL 100 for advanced writing, (c) English for native-like writing that is not academically competent, and (d) Portfolio for native-like advanced writing which meets the Foundation in Written Communication requirement of the UHM. The three graduate levels are (a) ELI 73 for intermediate writing, (b) ELI 83 for advanced writing, and (c) Exempt for writing that provides evidence that a student can meet or exceed expectations for the six above-mentioned determinants of *writing* proficiency [i.e., (a) to (f) in the previous section]. Previously,

the student essays on the ELIPT Writing Test were first rated by three raters separately based on a rubric describing necessary linguistic features for each of the course levels. Each rater would select the course level that they believed would suit that student. While rating student essays, raters were encouraged to go back to sample essays from each of the course levels to calibrate their placement decisions. If raters disagreed on placement decisions, the three raters discussed their decisions and tried to reach an agreement. When the three raters failed to reach an agreement, a fourth rater, usually a more experienced ELI instructor or administrator, joined the discussion and a final decision was made. The linguistic features of each course level included content, organization, vocabulary, grammar, and fluency (see Appendix A).

The raters were the ELI instructors, and they were not necessarily instructors of the writing courses. The ELI instructors (and raters) included bilingual, multilingual, and monolingual speakers of English, so they were not necessarily native speakers of English either. Some of them had learned English as a foreign language and demonstrated native-like proficiency in academic English. Most raters were experienced with the rating procedures, and when new raters joined the rating sessions, they participated in a training workshop held by the ELI associate director to familiarize them with the procedures, the descriptions in the rubric, the characteristics of each course level, and sample essays for each course level. Although there were five categories within each course level, only one holistic score, which was the course level, was given for each student essay. As a result, the five categories in the rubric may have been weighted differently by different raters and for different essays, causing concerns related to both rating reliability and validity. In fact, analyses of the writing test indicated low reliability for the previous rating rubric.

Revision Process

As the Executive Director of ESL Programs, Professor Brown proposed developing and using a new rubric based on English proficiency levels, instead of course levels. To those ends, he organized three workshops. The main purpose of the first workshop was to generate guidelines for developing a new rating rubric. Professors Brown and Kyle (serving as experts on language testing), the ELI director, the ELI associate director, the writing lead teacher, two ELI instructors, and a PhD student who has developed a rating rubric for the Hawaii English Language Program (HELP) participated the workshop. At the end of the first workshop, the

group decided on the following guidelines: a rubric should be developed that (a) is proficiency-based, (b) has individual score in each category, (c) has six points for each category across all the proficiency levels for both graduate and undergraduate students, (d) has five categories, which are the same as those in the previous rubric, and (e) uses category descriptions adapted from the previous rating rubric.

In the second workshop, the ELI director, associate director, and writing lead teacher finalized the category descriptions for each level and the category scores. In addition to the guidelines generated in the first workshop, during the discussion, it was suggested that the category descriptions from the previous rubric were ambiguous and had overlaps in the category descriptions. For example, the fluency category overlapped with the content category in that the amount of writing was closely related to the quality of the content in most student essays. Furthermore, the *grammar* category was relabeled to *grammatical complexity* to better reflect the focus on syntactical complexity rather than on individual grammatical errors. Another change to the new rubric was that the six-point scale for each category was changed to range from five to ten instead of the previous one to six. This change was used to reflect the full range of writing proficiency. At the end of the second workshop, it was decided that four, instead of five, categories should be used and the category descriptions and their scores were revised and more fully developed. After the second workshop, the new rubric was reviewed by Professor Brown, and the finalized rubric (see Appendix B) was used in the third workshop to investigate its reliability and validity.

In the third workshop, six ELI instructors participated; they re-rated all the student essays from the ELIPT administered for fall 2016. In total, 124 essays were numbered so raters would not know the test takers' identities. In the rating session, each essay was reviewed by two raters and two separate ratings were given to each essay.

Purpose of This Study

As mentioned above, the main purpose of the project was to revise a previous rating rubric used in the ELIPT at UHM. However, to evaluate the success or failure of such a project, data were gathered so that we could determine how well ratings based on the new rubric were spreading examinees out for purposes of making placement decisions into the writing courses in the ELI. In addition, it was important to know whether the resulting scores are reliable, how

many ratings are needed for each essay, and how well the subscales are performing. To those ends the following research questions were posed:

1. To what degree are the scores produced by the new rubric widely and normally distributed?
2. To what degree are the scores produced by the new rubric reliable/dependable?
3. To what degree are the two readings and four categories being rated severely/leniently?
4. To what degree are the six points in the four subscales contributing to the overall scores?

METHODS

Participants

Students. The 122 student essays[1] analyzed in this study were all written by international and immigrant students enrolled at the University of Hawai‘i at Mānoa, 41% were exchange undergraduate students, 38% non-exchange graduate students, 19% non-exchange undergraduate students, 2% exchange graduate students, and 1%, was a PBU student (1 student). In total, 66% were female and 34% male. In terms of previous proficiency scores, 9% of the examinees provided paper-based TOEFL scores, 68% internet-based TOEFL scores, and 17% IELTS scores. Table 1 is the summary of the examinee's scores on the standardized tests. The original placements (based on the earlier rubric) prior to this study were 42% into intermediate undergraduate writing (ELI 73), 25% into advanced undergraduate writing (ESL 100), 25% into advanced graduate writing (ELI 83), 3% exempted graduate writing, 2% native-like undergraduate writing (ENG 100), 2% undergraduate student who have fulfilled their writing requirements, and 1% between ESL 100 and ENG 100.

Table 1
Summary of the Student Essays' Proficiency Background

	pTOEFL	iTOEFL-WR	IELTS-WR
N	11	84	21
M	537.18	20.37	5.81
SD	16.88	2.77	0.45
Min	513	11	5
Max	580	26	6

[1] Two of the 124 original essays inadvertently received only one rating, so they were excluded from the subsequent analyses for a total of 122 essays.

Table 1 shows a summary of the participants' proficiency backgrounds in terms of proficiency scores that were previously available including the Paper-based TOEFL Test (pTOEFL), the Writing subtest of the Internet-based TOEFL Test (iTOEFL-WR), and the Writing subtest of the IELTS Test (IELTS-WR) including the number of participants involved, the mean (*M*), standard deviation (*SD*), minimum score (Min), and maximum score (Max).

Regarding the students' nationalities, approximately one-third (29%) of the student essays were written by Japanese, roughly one-third (21% + 7%) were written by Mainland Chinese and Taiwanese, and 15% were written by Koreans. A detailed list of all nationalities is provided in Table 2.

Table 2
Students' Nationality (in percentages)

Nationality	Percentage
Japan	29
China	21
South Korea	15
Taiwan	7
No Country Listed	6
Thailand	2
United States	2
Bhutan	2
Czech Republic	2
Philippines	2
Bangladesh	1
Brazil	1
Cuba	1
Denmark	1
Ecuador	1
French Polynesia	1
Hong Kong	1
Latvia	1
Mexico	1
Poland	1
Slovakia	1
Spain	1
Ukraine	1
Vietnam	1

Raters. The other participants in the project were six raters, who rated 124 writing samples written by the ELI Placement Test (ELIPT) examinees in fall 2016. The six raters were all ELI

instructors who were also graduate students, including four MA and two PhD students, enrolled in the Department of Second Language Studies. The background of the six raters is summarized in Table 3.

Table 3
Rater Backgrounds

Rater	Gender	English Proficiency	Class	ELI Teaching (in semesters)	Grading ELIPT-Writing (in times)
Rater 1	F	Native	MA	0	0
Rater 3	M	Native	MA	1	1
Rater 4	M	Native	MA	3	7
Rater 5	F	Native	MA	1	1
Rater 6	F	Native	PhD	1	2
Rater 8	F	Native-like	PhD	1	1

Materials

Existing system. When this project began, the existing system for writing placement (which had been used for 14 years) did not require raters to give a numerical score; rather, they would make a decision about which course would be most appropriate given the writing exhibited in a student's ELIPT Writing Test essay. To do this, they used a rubric that delineated hallmarks of typical writing at each course level for five categories of writing

Training involved review of the rubric and sample tests, which were selected as benchmarks for each level. Raters would compare the benchmark papers to the rubric, discuss questions and concerns, and thus, would calibrate themselves. By the time the training was done, the stack of tests would arrive from the test site.

For each test, there were three raters. After all the tests were read, the leader would call out a test-taker's name, and the three raters of that test would announce their recommended placements. If there was consensus among the raters, the placement would stand. When there was no consensus, the raters would discuss the strengths and weaknesses of that student's writing, and why a specific course seemed most appropriate, until agreement was reached (in cases of gridlock, the associate director, and sometimes the director, would make a final decision about placement).

The raters for the test included the associate director of the EAP program (who led the rating sessions) and graduate-assistant (GA) instructors in the program. Some of the GAs had two to

three semesters of teaching experience in the program, while others were about to begin their first semester in the program. A major concern about this system was that the GAs (particularly the new GAs) did not have enough familiarity with the program or its courses to be able to use the rubric effectively.

Professor Brown provided samples of different numeric rubrics used in other programs for rating academic writing, and one of the PhD students also provided a rubric that she had helped develop for the Hawai‘i English Language Program (HELP), an Intensive English Program at the pre-admission level, housed in Second Language Studies (the same department where the ELI was housed). Professor Brown highlighted the strengths of each rubric, and the participants in the meeting discussed the ELI's placement needs and what type of rubric might best serve those needs. One of the program's concerns was to balance efficiency (the ELI needed to rate writing tests within a few hours on the day the test was given, so that the program could provide test results to students the same day) and effectiveness (ratings that would be valid and consistently reliable). By the end of the first meeting, the group decided to adapt the wording of its hallmarks from the existing rubric to a new rubric that used a six-point scoring system for each of the five categories of writing from the existing rubric: content, organization, vocabulary, grammar, and fluency.

The ELI's director, associate director, and writing lead teacher had a second meeting to finalize the new rubric. One thing that became apparent quickly was that the category of *fluency*, as defined in the old rubric as the amount of writing generated in the 45-minute test, could be absorbed into the "content" category. The writing lead teacher suggested that the *organization* category be split into two new categories: *logical sequencing* and *cohesion*. After further discussion and efforts to tease out these two new categories, the group decided to abandon this division and include both ideas under the umbrella category of *organization*. Further, the category *grammar* was changed to *grammatical complexity*; even though it still addressed grammatical errors as well as complexity. However, the new rubric was designed to have complexity as the primary focus and errors as more of a secondary focus. After all of these changes, four categories remained: content, organization, vocabulary, and grammatical complexity.

One final change was an adjustment to the scoring system. The writing lead teacher suggested that most of the readers would feel that even the worst students' writing would not

warrant a score of 1, so the six-point system was adjusted to a range from 5 to 10, with an explanatory note that, "Because students in the ELI are at a level of academic English proficiency that is high enough to be admitted to UH Manoa, the 6-point numbering system used here ranges from 5-10, rather than 1-6." The ELI asked Professor Brown to review the rubric from his perspective as a testing expert, after which the rubric was finalized (Appendix B).

Procedures

Piloting the new rubric. To pilot the new rubric, the ELI arranged a four-hour block of time to re-rate writing tests from the fall 2016 ELI Placement Test. A total of 124 tests were re-rated and scored by two raters each using the new rubric independently (two essays were inadvertently rated by only one rater, so only 122 essays were analyzed in this study).

Nine GAs from the ELI originally signed up to re-rate the writing tests. Three did not show up for the reading, so only six raters participated in this study. Because some of the raters were teachers of ELI writing courses and would know any of the students who were enrolled in their courses, all tests were anonymized and randomly assigned a number between 1 and 124.

At the time the rubric was developed, it was suggested that no training would be necessary. Thus, for the re-rating pilot study, raters were asked to familiarize themselves with the rubric, and questions about interpretation of the descriptors were discussed.

Raters were given score sheets with the four categories across the top of each page, and the test-taker numbers in ascending order on the left side of each page. They were reminded to give a one-number score for each of the four categories for each test they read.[2] The re-reading of 124 tests was completed in approximately three hours.

Some of the raters mentioned that, perhaps because they were used to the previous system for placing students (which involved discussion whenever there was a discrepancy in placements assigned by different readers), reading and merely assigning scores felt more "mechanical and robotic."

The leader of the re-reading (the writing lead teacher) mentioned that he missed the discussion of student writing, which often involved discussion of how some of the projects and

[2] Despite this request, two of the readers apparently ignored this requirement, giving scores of a series of consecutive numbers (e.g., 5-6, or 7-8); when data was entered for these raters, the mean rating was used (e.g., 5.5, or 7.5).

writing assignments covered in ELI writing classes might address some of the test-takers' weaknesses. He felt that this was a lost "teacher training" opportunity that the previous system of scoring provided. The ELI director and associate director noted that perhaps this kind of discussion would need to be incorporated into new-teacher orientation, instead.

The writing lead teacher also mentioned that it felt like most of the raters were avoiding scores of 5-6 or 9-10, and that the majority of scores were falling within the 7-8 range. This suggests that perhaps the ELI needs to have rater training and develop a set of benchmark tests to use for calibrating raters across the entire range of scores.[3] However, it is also worth noting that in a normal distribution 68% (or a bit more than two-thirds) of the scores are expected to cluster around the mean (i.e., fall within one standard deviation of the mean).

RESULTS

Descriptive Statistics

Table 4 presents the descriptive statistics for the six raters and 122 examinees in this study. Six statistics are listed in the first column (*N* = the *number* of essays read by each rater; the *mean* in this case is equivalent to the arithmetic average; the *median* is that point the divides the scores 50/50; the *SD* = the *standard deviation* is an indicator of the dispersion of the scores because it is a sort of average of the distances of scores from the mean; the *min* is the minimum or lowest score given; the *max* is the maximum or highest score given; the *range* is another indicator of the dispersion of scores because it represents the distance between the highest and lowest score including both of them; the *skew* statistic is an indicator of the degree of skewedness or non-normality, when it is interpreted in light of the *SES* (or standard error of skew); and the Cronbach alpha (α) is an estimate of the reliability of the scores generated by each rater across the four categories. In addition, the six raters are labeled across the top (Raters 1, 3, 4, 5, 6, & 8) along with the two-reading average for all scores (i.e., the average of whichever two raters scored each essay) over to the right.

[3] Unfortunately, at the time of this study, the ELI did not have the time or resources to develop a set of benchmark tests, or a training plan.

Table 4
Descriptive Statistics for Six Raters and Two-Rater Averages

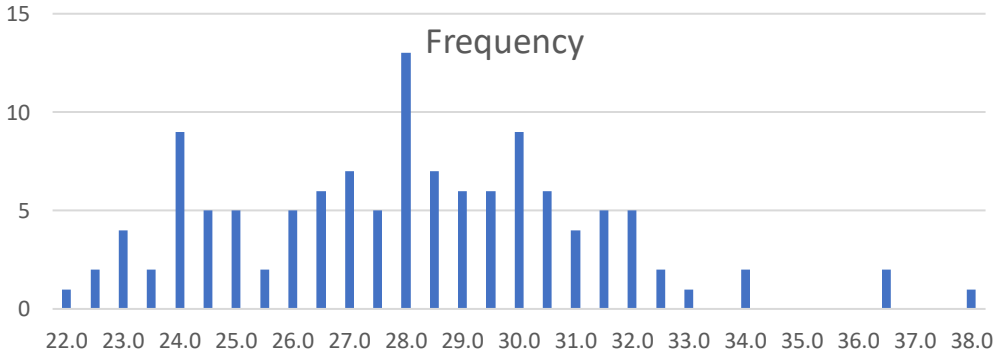
Statistic	Rater 1	Rater 3	Rater 4	Rater 5	Rater 6	Rater 8	Two-reading Average
N	39	46	43	35	34	47	122
Mean	26.36	28.72	27.12	26.09	30.59	29.28	28.05
Median	25.0	27.5	26.0	25.0	30.0	30.0	28.0
SD	4.44	3.76	3.13	2.75	2.45	4.07	3.13
Min	20	24	22	23	24	22	22
Max	38	39	35	34	38	38	38
Range	19	16	14	12	15	17	17
Skew	0.72	0.67	0.52	1.20	0.10	-0.13	0.37
SES	0.39	0.36	0.37	0.41	0.42	0.36	0.22
Cronbach α	.904	.883	.902	.886	.634	.915	.89[a]/.90[b]

a. For the first reading
b. For the second reading

Clearly Rater 6 had the highest average scores with a mean of 30.59 and Rater 5 had the lowest with 26.09. The standard deviations indicate that Rater 1 had the scores that varied the most with an *SD* of 4.44, while Rater 6 had the scores that varied the least with an *SD* of 2.45. The ranges also indicate that Rater 1 had the scores that varied the most with a range of 19, but that Rater 5 had the scores that varied the least with a range of only 12.

The column furthest to the right in Table 4 analyzes the overall the scores generated in this study ($N = 122$) from averaging the two readings that each essay received. These overall scores were reasonably well-centered with a mean of 28.05, which is very nearly the same as the median of 28.0. These scores were also well dispersed with a standard deviation of 3.13 and range of 17 with scores ranging from a low of 22 to a high of 38. The Skew statistic and standard error of skew (SES) indicate that the distributions of scores were approximately normal for all raters scores, except for Rater 5, for whom the skew statistic is 1.20, which is more than two times the SES value of 0.41 in a positive direction indicating positive skew. Figure 1 provides a visual representation of the distribution of two-reading average scores. This distribution looks approximately normal, except for the three scores over to the right, which may be outliers. The Cronbach alpha estimates will be discussed in the next section.

Figure 1. Two-Reading Average Writing Scores Labeled Horizontally and Frequencies Labeled Vertically



Because the Rasch analyses discussed below examines the effects of differences in severity/leniency for the first and second readings of the essays as well as variations in the relative ratings for the four categories in the rubric, Table 5 presents the descriptive statistics for the four categories (Content, Organization, Vocabulary, and Grammatical Complexity) and Total scores (in bold type) for Reading 1 (R1) and Reading 2 (R2). Notice that the overall mean for R2 is 1.46 higher (28.78 – 27.32 = 1.46) and the median is 4 points higher (30 – 26 = 4) than those for R1. Notice also that the category means for R1 range narrowly from 6.74 to 7.07 and that the medians are all the same 7.0, while the means for the categories in R2 range from 7.14 to 7.24 and the medians are all 7.5. The moderately higher scores for R2 probably resulted from the fact that two raters in that second reading unaccountably decided to give double scores (e.g., 7/8), which we coded as halfway between (e.g., 7.5), while none of the raters in the first reading did so.

Table 5
Descriptive Statistics for Two-Reading Totals and Categories

Statistic	R1 Cont	R1 Orgn	R1 Vocb	R1 Gram	**R1 Total**	R2 Cont	R2 Orgn	R2 Vocb	R2 Gram	**R2 Total**
N	122	122	122	122	**122**	122	122	122	122	**122**
Mean	7.07	6.74	6.75	6.76	**27.32**	7.23	7.17	7.14	7.24	**28.78**
Median	7.0	7.0	7.0	7.0	**26.0**	7.5	7.5	7.5	7.5	**30.0**
SD	1.20	1.06	0.99	1.00	**3.70**	1.05	1.21	1.08	1.15	**3.93**
Min	5	5	5	5	**20**	5	5	5	5	**22**
Max	10	10	9	9	38	10	9.5	10	10	**39**

Classical Theory Reliability Estimates

The Cronbach α values shown at the bottom of Table 4 indicate the reliability of scores across the four categories for each of the raters. Reliability coefficients like these can be interpreted as the proportion of consistent variance found in each set of scores. For example, Rater 1 had a Cronbach α value of .904, so it can be said that about 9/10s of the variance in this rater's category scores was consistent, while 1/10 was not, or more precisely (by moving the decimal point two places to the right) that 90.4% of the variance in this rater's category scores was consistent, while 9.6% was not. Notice that the scores are acceptably reliable for raters 1, 3, 4, 5, and 8, with values ranging from .883 to .915. The remaining rater was not quite so consistent with a Cronbach α of .634, which means that only 63.4 percent of the variance in this rater's category scores was consistent while 36.6% was not. Such internally inconsistent raters may benefit from a clearer rubric, additional training in using the rubric, or added experience in doing such ratings. Note that the Cronbach α estimates in the last column of Table 4 (across categories for the first and second readings of these essays) were .89 and .90, respectively; these indicate the single-reading reliabilities. Adjusted for two-readings, which would naturally be more reliable than one reading because there are more observations (using the Spearman-Brown prophecy formula), the two-reading reliabilities were.94 and .95, respectively.

Table 6 shows the correlations (*r*) and numbers (*n*) of pairs involved in each case for Raters 1, 3, 4, 5, 6, and 8; the correlation coefficients range from .04 to .98 in the positive direction and in one case as high as -.67 in the negative direction with as few as 0, 1, 3, or 4 pairs involved and as many as 16 pairs involved. This wide and inconsistent range of correlation coefficients indicates a need for studying reliability more systematically by having raters work in pairs on the same essays in the future. Because these correlation coefficients are based on non-systematic pairings of raters, they are difficult to interpret. In the future, such studies should probably be set up to carefully pair raters so that each composition is read by two raters, and each rater pair reads about the same number of essays.

Table 6
Correlation Coefficients Between Pairings (non-systematic)

		Rater 1	Rater 3	Rater 4	Rater 5	Rater 6	Rater 8
Rater 1	*r*	1.00	0.74	*	*	0.66	0.41
	n	39	14	1	1	8	15
Rater 3	*r*		1.00	0.44	0.60	*	0.98
	n		46	16	13	0	3
Rater 4	*r*			1.00	*	.04	0.61
	n			43	0	15	11
Rater 5	*r*				1.00	0.50	0.27
	n				35	7	14
Rater 6	*r*					1.00	-0.67
	n					34	4
Rater 8	*r*						1.00
	n						47

* Cannot be computed with no variance

Table 7 shows the correlations coefficients between the ratings for each category and all other categories for the two-reading paired scores. The two readings were done for all examinees so there are 122 examinees used for each correlation coefficient. The category-total correlations are shown in bold type, and these correlations are naturally higher because they are inflated by the fact that the category involved in each case is part of the total. However, the correlation between the totals for reading 1 and 2 (shown in the column furthest to the right in bold italics) is only .346. This value of .346 represents the reliability of either reading 1 or reading 2 but not the two taken together. Using the Spearman-Brown adjustment, the reliability of two readings would be .51, and adjusted for three readings, it would be .61. This is one indication that at least three readings *may* be necessary if the ELI wants to obtain inter-reading reliability of even 61% for the writing scores on the ELIPT.

Table 7
Correlation Coefficients Between Paired Readings (systematic with N *= 122 in all cases)*

	R1 Cont	R1 Orgn	R1 Vocb	R1 Gram	**R1 Total**	R2 Cont	R2 Orgn	R2 Vocb	R2 Gram	**R2 Total**
R1 Cont	1.000	0.745	0.643	0.612	**0.875**	0.375	0.220	0.242	0.222	0.299
R1 Orgn		1.000	0.686	0.641	**0.885**	0.375	0.197	0.267	0.217	0.297
R1 Vocb			1.000	0.735	**0.871**	0.284	0.069	0.264	0.198	0.227
R1 Gram				1.000	**0.849**	0.385	0.253	0.349	0.362	0.382
R1 Total					1.000	0.409	0.215	0.320	0.285	***0.346***
R2 Cont						1.000	0.680	0.677	0.637	**0.848**
R2 Orgn							1.000	0.697	0.683	**0.881**
R2 Vocb								1.000	0.761	**0.892**
R2 Gram									1.000	**0.881**
R2 Total										1.000

Correlations Between the Two Rubrics

To calculate the degree of relationship between the old and new rubric scores, two correlation coefficients were calculated here: Spearman's rho and Goodman and Kruskal's Gamma coefficient. The first variable was based on the ratings generated for these essays using the original rubric. Essays which were originally rated as ELI 73 were converted to intermediate (or 1), ESL 100 and ELI 83 were considered advanced (or 2), and ENG, Portfolio, and Exempt was exempt (or 3). The second variable represents the new placement decisions generated from new rubric and based on the new cut-scores (see Conclusion section below).

We first calculated Spearman rho (ρ) to investigate the degree of relationship if these placement levels are treated as ordinal scales, and then calculated Goodman and Kruskal's Gamma because it is more suitable when there are limited levels in the ordinal scale. Spearman's rho turned out to be an astounding 0.9998, which is misleading due to the relatively small degree of disagreement caused by only three levels in the two variables in comparisons to 122 test takers. In contrast, the Goodman and Kruskal's Gamma coefficient was 0.61 ($p < 0.01$) better reflects the (mis)match between the two variables.

Generalizability Theory Dependability Estimates

Generalizability theory (G theory) is an extension of classical testing reliability theory, which first appeared in the literature in Cronbach, Rajaratnam, and Gleser (1963). G-theory uses intermediary mean squares figures from analysis of variance (ANOVA) procedures to identify

and estimate the relative amounts of variance (called variance components, or VCs) for various measurement *facets* as identified by the testing researcher. VCs are important because they can be used to study the relative effects of increasing or decreasing the numbers of units in each facet on score *dependability* (analogous to *reliability*). Decisions about the best numbers of raters and/or items rating to use to get the desired level of dependability given the practical constraints in the testing situation. [For much more on G-theory, see Brennan (2001); Chiu (2001); McNamara (1996); and Shavelson & Webb (1991).]

G-study variance components. Generalizability research is usually carried out in two stages. The first stage is called the generalizability study (or simply G study). *Variance components* (VCs) are derived for each of the facets and all possible interactions of facets using ANOVA procedures. For example, in the current ELIPT Writing Test study, we were interested in the relative effects of three facets, that is, persons (122 individual examinees), readings (1st and 2nd) and categories (content, organization, grammatical complexity, and vocabulary) in a persons by readings by categories design (p x c x r), and all possible interactions (pc, pr, rc, pcr). Using the GENOVA program, which is downloadable at (http://www.education.uiowa.edu/casma/GenovaPrograms.htm) free of charge, we calculated the p x c x r ANOVA results as shown in first four columns of Table 8. We also calculated variance components and their standard errors as shown in Table 9.

Table 8
ANOVA for p x c x r Design

Source	*df*	*SS*	*MS*
Persons	121	578.6967	4.78262
Readings	1	1.04918	1.04918
Categories	3	6.60246	2.20082
pr	121	251.2008	2.07604
pc	363	160.6475	0.44256
Rc	3	3.90164	1.30055
prc	363	120.8484	0.33292

Table 9
Variance Components, Percentages of Variance, and Standard Errors

Effect	*df*	VCs (EMS Equations)	Percentages of Variance	Standard Errors
Persons	121	0.3246173	28.00%	0.0832622
Readings	1	0.0000000	0.00%	0.0024939
Categories	3	0.0032403	0.28%	0.0066283
pr	121	0.4357811	37.59%	0.0664679
pc	363	0.0548198	4.73%	0.0204968
rc	3	0.0079314	0.68%	0.0067451
prc	363	0.3329156	28.72%	0.0246435
Total Variance		1.1593055	100.00%	

The clearest way to explain the VCs for each facet and their interactions is to examine their relative magnitudes by converting the VCs into percentages of the total variance as shown in the fourth column of Table 9. The VC for persons is fairly large (at 28%), which is as it should be in a norm-referenced test like the ELIPT Writing Test. Readings, categories, and their interaction (rc) contributed only small percentages of variance, with 0.00%, 0.28%, and 0.68%, respectively, while the two-way interaction for pc accounted for more variance with 4.73%, the remaining interactions for pr and prc accounted for substantial variance with 37.59% and 28.72%, respectively. What this means is that the persons VC is spreading people out reasonably well as would be expected, but the pr, pc, and prc interactions are showing the extent to which raters and categories are inconsistent across persons, and each other. Thus finishes the first G-study stage.

Decision-study what-if estimates. The second stage in G-theory research is called the decision study (D study). In this stage, the G-study VCs are used to calculate dependability estimates, which are analogous to classical test theory reliability coefficients, for various testing situations. In this p x r x c study, we examine the dependability estimates for various numbers of raters and categories with the goal of understanding various possible test designs to see which will be most dependable in a revised version of the ELIPT Writing Test in terms of dependability given practical constraints on numbers of raters and numbers of categories.

Table 10
Generalizability Coefficients (for Relative Decisions, NRT) with Different Numbers of Categories and Raters

		Categories														
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Readings	**1**	.28	.34	.36	.38	.39	.39	.40	.40	.40	.41	.41	.41	.41	.41	.41
	2	.43	.50	.53	.54	.55	.56	.57	.57	.57	.57	.58	.58	.58	.58	.58
	3	.51	.59	.62	***.63***	.65	.65	.66	.66	.66	.67	.67	.67	.67	.67	.67
	4	.57	.65	.68	.69	.70	.71	.72	.72	.72	.73	.73	.73	.73	.73	.73
	5	.61	.69	.72	.73	.74	.75	.76	.76	.76	.77	.77	.77	.77	.77	.77
	6	.64	.72	.75	.76	.77	.78	.79	.79	.79	.80	.80	.80	.80	.80	.80
	7	.66	.74	.77	.79	.80	.80	.81	.81	.82	.82	.82	.82	.82	.82	.82
	8	.68	.76	.79	.81	.81	.82	.83	.83	.83	.84	.84	.84	.84	.84	.84
	9	.70	.77	.80	.82	.83	.84	.84	.84	.85	.85	.85	.85	.85	.86	.86
	10	.71	.79	.82	.83	.84	.85	.85	.86	.86	.86	.86	.86	.87	.87	.87
	11	.72	.80	.83	.84	.85	.86	.86	.87	.87	.87	.87	.87	.88	.88	.88
	12	.73	.81	.84	.85	.86	.87	.87	.87	.88	.88	.88	.88	.88	.88	.89
	13	.74	.81	.84	.86	.87	.87	.88	.88	.88	.89	.89	.89	.89	.89	.89
	14	.75	.82	.85	.86	.87	.88	.88	.89	.89	.89	.89	.90	.90	.90	.90
	15	.75	.83	.86	.87	.88	.89	.89	.89	.90	.90	.90	.90	.90	.90	.90

The values in Table 10 were calculated for varying numbers of readings and categories. Notice in Table 10 that the numbers of categories are labeled across the top and the numbers of readings down the left side of the table. Thus, the table shows what is likely to result if different numbers of categories or readings are used in the future. For example, if four categories were used and three raters, and the resulting dependability estimate would be .63 (see bold italics in Table 10). In the present study, we had four categories and two readings per essay. Table 10 indicates that the dependability under these conditions was .54. The table also shows that if six categories were used with three readings, the dependability would only increase by a little to .65. Clearly, Table 10 indicates what the dependability increases more by adding readings than it does by adding categories.

Multifaceted Rasch Analysis

We used multifaceted Rasch analysis (based on the FACETS program, see Linacre, 2010) in this study to simultaneously examine the degree to different levels of variables produce distinct scores relative to each other—all on the same scale called a logit scale. In this study, we were of course interested in how well the scoring procedures that we used spread the examinees out, but

also the degree to which different raters were severe or lenient and rating categories were difficult or easy.

Model fit. Table 11 summarizes the results for the Facets analysis conducted in this study. Notice in the first column that the rows are labeled with the three facets (Examinees, Readings, and Categories). Also note that labels are given across the top for five statistics: Number of Misfits, RMSE, Separation Index, Reliability, and Chi-square (fixed), each of which will now be considered in turn.

Table 11
Facets Analysis Statistics

Facet	Number of Misfits	RMSE	Separation Index	Reliability	Chi-square (fixed)
Examinees	4	.49	2.76	.88	$p = .00$
Readings	0	.06	.96	.48	$p = .17$
Categories	0	.09	2.09	.81	$p = .00$

The *Number of Misfits* indicates how many examinees, readings, or categories "did not fit the general pattern of responses in the matrix, and can thus be classified as relatively misfitting…" (McNamara, 1996, p. 171). Table 11 shows that there were four misfitting examinees and that there were no misfits for readings or categories, which means that four examinees were not fitting the measurement model in this analysis due to score patterns that were not expected. Note that none of the examinees received perfect scores of 40, which in Rasch analysis means that the test was appropriate for all of these examinees because it could estimate scores within the scale.

RMSE is the *root mean square standard error*, which is an intermediary result in calculating the separation index discussed next. However, RMSE itself provides an estimate of standard error. Thus the lower the *RMSE* is, the better the data are fitting the measurement model. The *RMSE* values in Table 11 for readings and categories are .06 and .09 respectively, which indicates reasonably good fit to the measurement model. The same cannot be said for the RMSE of .49 for Examinees, which is indicating that examinees are not fitting the model as well as might be desired. This may be caused by the four misfitting examinees.

The *separation* index indicates the degree to which the three facets (examinees, readings, and categories) are spread out relative to their precision of measurement (Linacre, 2010, p. 160). The higher the value, the more a facet is spreading its components out. Notice that the separation indexes for examinees and categories are higher than the one for other readings, which is

relatively low. All of this indicates that the examinees and categories facets are providing better estimate spreads relative to measurement precision than readings are. This makes sense given that that mean logit scores for examinees and categories vary considerably from each other while those for the two readings are nearly the same.

The *reliability* values reported in Table 11 might more exactly be called *separation reliabilities*. For example, the reliability of .88 for examinees means that the examinees consistently varied from each other, which is typically viewed as a good thing in norm-referenced testing. In comparison, the relatively low reliability for raters of .48 would typically be viewed as "desireable" because it indicates that the raters are *not* very consistently different from each other in terms of the leniency or severity of the ratings they give. The degree of reliability among categories is different yet again. From the point of view of the ELIPT Writing Test, it is not a problem if one category is consistently scored lower or higher than the other categories. So, the consistent differences among categories indicated by the separation reliability of .81 pose no problem. In short, these reliability estimates appear to indicate that the three facets are operating as they should with reasonable and appropriate consistency.

The *chi-square (fixed)* statistics in this study test the following three propositions:

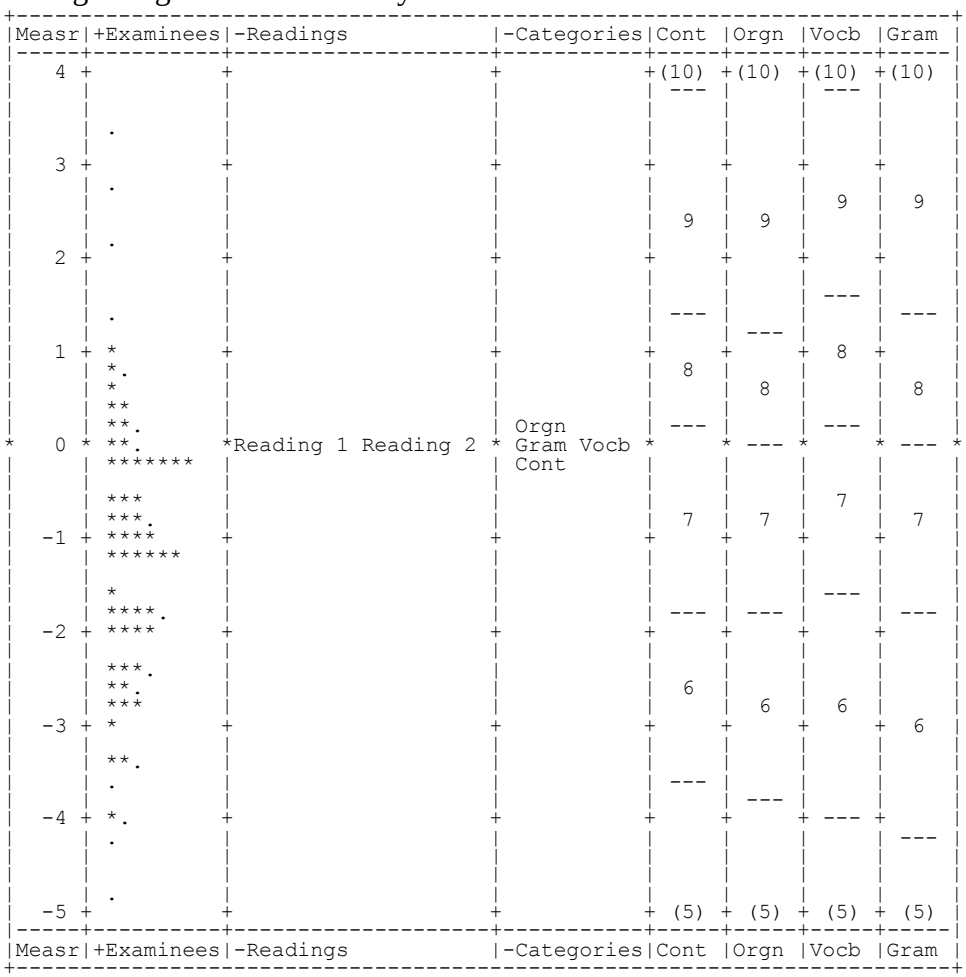
1. These examinees can be considered equally able
2. These readings be considered equally severe or lenient
3. These categories considered equally difficult

The chi-square statistics in this study were found to be significant (at $p < .01$) for examinees and categories, but not for readings. Thus propositions 1 and 3 can be rejected, while proposition 2 must be accepted, that is, the examinees cannot be considered equally able and the categories cannot be considered equally difficult, but the readings can be considered equally severe or lenient.

Vertical ruler. Next, we will interpret the vertical ruler from our FACETS analyses. Notice in Figure 2 that the first column in the vertical ruler is for measure (Measr), which shows the full range of scores in logits (a true interval logit scale where the mean is 0 and, in this case, the possible range is labeled from almost -5 to +4. The second column shows where the examinees logit scores were on the scale (with each asterisk equivalent to 2 examinees and a dot equal to one). The third column gives the logit averages for each of the two readings (Reading 1 and Reading 2). The fourth column shows that the logit average ratings for each of the four rating

categories: Content (Cont), Organization (Orgn), Vocabulary (Vocb), Grammatical Complexity (Gram). The final four columns show the distributions of the 5 to 10 scores for each of the categories. The plus and minus signs in the headings (i.e., for +Examinees, -Readings, and -Categories) indicate the direction of the scale. Examinees are plus and arranged form high ability examinees at the top to low ability students at the bottom; readings are minus and arranged from severe ratings at the top and lenient ones at the bottom; and categories are also minus and arranged from severe ratings at the top and lenient ones at the bottom.

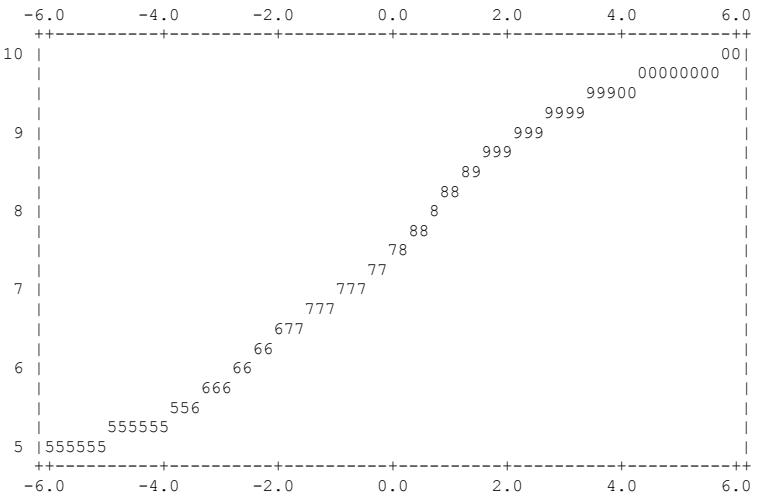
Figure 2. Partial Credit Model Vertical Ruler for the 122 Examinees, Two Readings, and Four Rating Categories in This Study



Notice in Figure 2 that the logit scores of the examinees ranged from almost -5 to a bit above +3 with many scores clustered symmetrically around -1. In more exact terms, the logit mean for the 122 examinees was -1.18. Note also that the two readings do not differ enough to show any

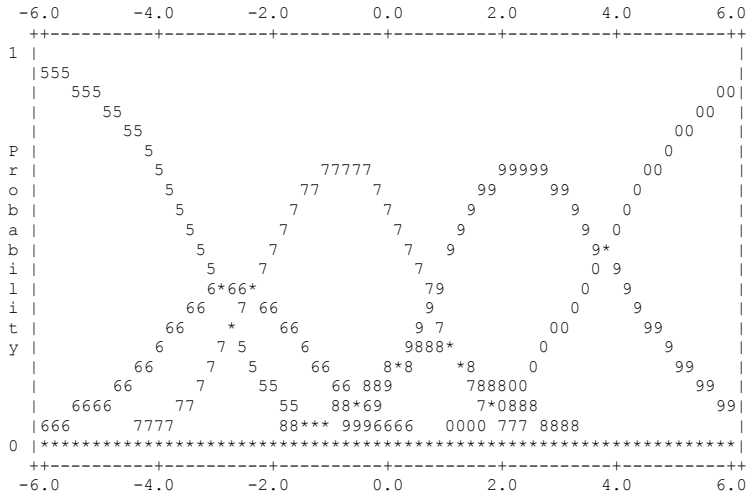
difference in Figure 2 (more precisely, the logit means for the readings 1 and 2 were +.06 and -.06, respectively). However, categories differ visibly with organization being the most severely rated; content the most leniently rated; and grammatical complexity and vocabulary about equally severe/lenient in between the other two (more precisely, the logit means for organization, grammatical complexity, vocabulary, and content were +.19, +.06, +.03, and -.28, respectively). The final four columns illustrate how the raw scores (from 5 to 10) for content, organization, vocabulary, and grammatical complexity were distributed along the same scale relative to logits, examinees, readings, and categories.

Figure 3. Expected Score Ogive Cure (Model ICC)



The Expected score ogive curve for the score values from 5 to 10 (with 0 representing 10) are shown in Figure 3. This curve indicates the model curve predicted from these data for the overall the writing test scores. In general, the discrimination is better at those points of the curve where the angle is steeper (i.e., in the range between 6 and 8).

Figure 4. Probability Curves for Each Score Value (with 0 representing 10)



The probability curves for each of the scores values (with the 0s representing 10) are shown in Figure 4. Generally, the different score values are performing better if they appear in such a graph as distinct mounds or hills. Thus, 5, 7, 9, and 10 appear to be functioning well, but 6 and 8 do not appear to be distinct from the categories on either side of them. This analysis suggests that the raters may still be matching their scores with course levels in their minds, leading to them avoid scores between course levels. Given that, it might be better to abandon the six-point scale, and use a four-point scale instead. Perhaps the rubric rows could be changed to represent scores of 7, 8, 9-10 and the categories could be labeled with scores ranging from 7-10.

DISCUSSION

In order to help readers sort through all the results in the previous section, we will answer the four research questions posed at the beginning of this study in the clearest possible prose without using technical jargon unless absolutely necessary. In addition, to make the organization of the discussion clear, we will use the research questions themselves as headings.

To what degree are the scores produced by the new rubric widely and normally distributed?

The descriptive statistics in this study (see Tables 4 & 5 and Figure 1) indicate that the total scores produced by the ELIPT Writing Test using the new rubric are sufficiently spread out and reasonably normal in distribution.

To what degree are the scores produced by the new rubric reliable/dependable?

The Cronbach α reliability estimates produced by the individual raters indicate that their ratings were 90.4%, 88.3%, 90.2%, 88.6%, 63.4%, and 91.5% reliable (see Table 4). Furthermore, the Cronbach α estimates across categories for the first and second readings of these essays indicate that the ratings were 89% and 90% reliable, respectively. The interrater correlations show in Table 6 were of little use because the ratings were not done systematically in rater pairs or trios. However, the inter-reading reliability between the totals for reading 1 and 2 (shown in Table 7) turned out to be about 35%, but adjusted for two readings the inter-reading reliability was 51%, and adjusted for three readings, it was 61%.

The first stage of the G-theory study indicates that there are considerable pr, pc, and prc interactions, which mean that the ratings for readings and categories are inconsistent across persons and each other. The second stage indicates that the dependability of the ELIPT Writing Test would be increased more by changing (from two to three) the number of readings than by changing the number of categories.

The multi-faceted Rasch analysis produced a separation reliability estimate of .88 for examinees, which is roughly analogous with Cronbach alpha. The .41 separation reliability for readings is a positive development because we do not want readings to be systematically different, while the .81 reliability for categories is fine because it indicates that the categories are being applied differently but with reasonable consistency. These Rasch results are true for the logit scores produced by the Rasch analysis, but not necessarily for the raw scores upon which those logit scores are based. In other words, the logit scores form a true interval scale that corrects for differences in readings and categories, all of which may not be true for the raw scores.

To what degree are the two readings and four categories being rated severely/leniently?

The multifaceted Rasch analysis vertical rulers (Figure 2) further indicated that the logit scores for examinees were reasonably well distributed from a low of almost -5 to a high a bit above +3 with many scores clustered symmetrically around -1. While the two readings did not differ much, the categories differed more with organization being the most severely rated, content being the most leniently rated, and grammatical complexity and vocabulary being about equally severe/lenient in between the other two.

To what degree are the six points in the four subscales contributing to the overall scores?

The probability curves shown in Figure 4 indicate that the scale points 5, 7, 9, and 10 are functioning well, but 6 and 8 are not distinct from the scale points on either side of them. Thus, instead of a six-point scale, a four-point scale might be more appropriate—perhaps with the rubric rows represented as scores of 7, 8, 9-10 and the categories labeled with scores ranging from 7-10.

CONCLUSIONS

Limitations

The present study had at least four limitations. First, little or no training was done for the raters because we wanted to see if the new rubric would work well without training. Second, as indicated in Table 6, we did not gather the ratings in a systematic way that would have allowed studying rater pairs or trios for consistency (i.e., interrater reliability and generalizability theory). Third, some of the raters unexpectedly used two scores which we coded as .5 scores that fell between their two points on the scale. Fourth, the sample size of 122 essays, while adequate could have been bigger. We will next suggest ways to overcome these limitations in the next section along with other ideas for improving the ELIPT Writing Test rubric and procedures in the future.

Suggestions for Improving the ELIPT Writing Test

The following suggestions are based on the findings and experience gained during this research project:

1. Raters should be trained by going over the rubric so that everyone understands how it works (including the fact that no double or .5 scores should be assigned). That training should also afford raters the opportunity to rate some sample essays at the different levels on the rubric. Such practice need not be extensive, but it does need to familiarize the raters with the entire range of possible levels from just below the SLS 73 level to Exempt level essays.
2. During the actual rating process three raters should probably be used for each essay.

3. Raters should work in trios, such that all three raters rate the same essays—preferably in a balanced design with roughly the same number of essays assigned to each trio.
4. Raters should work independently in rating each essay.

Cut-scores for the ELIPT Writing Test

Based on the scoring of the new rubric, the possible highest and the lowest scores are 40 and 20, respectively. All the essays were firstly placed into three levels based on the total scores from the four categories. Intermediate essays are those scored at 28 or below; advanced essays are those scored in the range from 29 to 31; exempt essays are those scored equal to 32 or higher. The corresponding course levels to the three levels were listed in Table 12 for undergraduate (UG) and graduate students (G).

Table 12
Cut-Scores of Course Levels

	Cut-Scores	Course Levels
Exempt	32 or higher	ENG (for UG) Portfolio (for UG) Exempt (for G)
Advanced	29 to 31	ESL 100 (for G) ELI 83 (for G)
Intermediate	28 or lower	ELI 73 (for UG)

REFERENCES

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

Chiu, C. W.-T. (2001). *Scoring performance assessments based on judgments: Generalizability theory*. Boston: Kluwer Academic.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology, 16,* 137-163.

Linacre, J. M. (2010). *A user's guide to FACETS Rasch-model computer programs: Program manual 3.67.0.* www.winsteps.com.

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

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage.

ACKNOWLEDGMENTS

The authors would like to thank the following people in the Department of Second Language Studies (SLS) for their support, which was invaluable to this project. First, we appreciate the work of the six graduate assistants in the ELI—Kelly Bolen, Lucas Edmond, Uy-Di Nancy Le, Anna Mendoza, and Lin Zhou—who served as raters for the re-rated of writing samples, as well as Daniel Holden (writing lead teacher in the ELI), who was instrumental in helping develop the new ELI rubric and leading the re-rating session. Additional thanks go to Dr. Kristopher Kyle (graduate faculty member in SLS) and to Patharaorn Patharakorn (PhD student in SLS) for sharing their insights during discussion of rubrics. Further thanks to Kristen Rock (PhD student in SLS) for sharing the rubric she created for rating the writing tests in the Hawai`i English Language Program), as well as for her insights during the discussion of rubrics. Particular thanks go to Priscilla Faucette (Associate Director of the ELI) for her wisdom, insights, patience, and support of the new rubric and rating system.

Appendix A: Original Rubrics for ELIPT Writing Test

Graduate Students

	Content	**Organization**	**Vocabulary**	**Grammar**	**Fluency**
Exempt *Shows high proficiency in L2, knowledge and control of academic writing genres/ conventions common in US universities.*	*Paper shows evidence of:* • Clear point/argument • Complexity of thought/ analysis • Insight on the topic, rather than mere description • Ample supporting evidence, detail, and examples • Command of rhetoric, evidence of writing style	*Paper is:* • Cohesive • Well-developed • Unique (not formulaic) • Marked by clear transitions, appropriate use of transitional phrases	*Paper has:* • Wide variety of vocabulary • Appropriate use of idioms • Few problems with collocations • Few problems with word choice	*Paper has:* • Some errors, but none that interfere with comprehension • Complex sentence structure (e.g., complex coordination, subordination, embedded questions, etc.)	*Amount of writing is:* • Suitable for level of analysis and/or amount of time provided to write the paper
ELI 83 *Shows some knowledge and control of academic writing; needs to develop L2 proficiency, writing ability, and/or awareness of genres/conventions common in US universities.*	*Paper shows evidence of:* • Clear, developed argument, but may be simplistic • Some insight on the topic, but may lack depth • Effective support, but evidence and examples may be general or vague	*Paper is*: • Cohesive • Somewhat formulaic (e.g. 5-para essay format) • Marked by appropriate transitions, with some misuse/overuse of transitional phrases	*Paper has:* • Varied vocabulary • Some problems with collocations • Some problems with word choice	*Paper has:* • Several errors (e.g., tense/aspect, word form, articles, pre-positions), but typically do not interfere with comprehension • Some correct complex sentence structure; evidence of other (incorrect) attempts	*Amount of writing is:* • Suitable for level of analysis and/or amount of time provided to write the paper
ELI 73 *Needs to develop L2 proficiency; notable unfamiliarity with and general lack of control of academic writing; would benefit from at least two semesters of ELI writing instruction.*	*Paper shows evidence of:* • Undeveloped or unclear argument • Simple topic description/ restatement, but with little insight • A general lack of supporting evidence, detail, examples • Redundancy of ideas, argumentation	*Paper is:* • Not cohesive • Formulaic (e.g., 5-para essay format), or lacking organization • Marked by absence of clear transitions between ideas, or simple sentence-level transitions used at paragraph level (e.g., first, next, then)	*Paper has:* • Notably limited vocabulary • Repetition/overuse of certain lexical items • Numerous problems with word choice • Incorrect collocations	*Paper has:* • Numerous errors that typically interfere with comprehension • General lack of sentence complexity	*Amount of writing is:* • Unsuitable for level of analysis and/or amount of time provided to write paper

Undergraduate Students

	Content	**Organization**	**Vocabulary**	**Grammar**	**Fluency**
English *Shows high proficiency in L2, but need for instruction in rhetoric, organization, support, argumentation. Will benefit from English Dept rather than ELI writing instruction.*	*Paper shows evidence of:* • Clear, developed argument, but may be simplistic • Some insight on the topic, but may lack depth • Effective support, but evidence and examples may be general or vague	*Paper is*: • Cohesive • Somewhat formulaic (e.g. 5-para essay format) • Marked by appropriate transitions	*Paper has:* • Varied vocabulary • Few problems with collocations • Few problems with word choice	*Paper has:* • Few errors • Complex sentence structure (e.g., complex coordination, subordination, embedded questions, etc.)	*Amount of writing is:* • Suitable for level of analysis and/or amount of time provided to write the paper
Portfolio *Shows high proficiency in L2, knowledge and control of academic writing genres/ conventions common in US universities. May not benefit from writing instruction.*	*Paper shows evidence of:* • Clear point/argument • Complexity of thought/ analysis • Insight on the topic, rather than mere description • Ample supporting evidence, detail, and examples • Command of rhetoric, evidence of writing style	*Paper is:* • Cohesive • Well-developed • Unique (not formulaic) • Marked by clear transitions, appropriate use of transitional phrases	*Paper has:* • Wide variety of vocabulary • Appropriate use of idioms • Few problems with collocations • Few problems with word choice	*Paper has:* • Few errors • Complex sentence structure (e.g., complex coordination, subordination, embedded questions, etc.)	*Amount of writing is:* • Suitable for level of analysis and/or amount of time provided to write the paper
ELI 100 *Shows some knowledge and control of academic writing; needs to develop L2 proficiency, writing ability, awareness of genres/conventions common in US universities. Will benefit from ELI rather than English Dept instruction.*	*Paper shows evidence of:* • Clear, developed argument, but may be simplistic • Some insight on the topic, but may lack depth • Effective support, but evidence and examples may be general or vague	*Paper is*: • Cohesive • Somewhat formulaic (e.g. 5-para essay format) • Marked by appropriate transitions, with some misuse/overuse of transitional phrases	*Paper has:* • Varied vocabulary • Some problems with collocations • Some problems with word choice	*Paper has:* • Several errors (e.g., verb tense/aspect, word form, articles, pre-positions), but typically do not interfere with comprehension • Some correct complex sentence structure; evidence of other (incorrect) attempts	*Amount of writing is:* • Suitable for level of analysis and/or amount of time provided to write the paper
ELI 73 *Needs to develop L2 proficiency; notable unfamiliarity with and general lack of control of academic writing; would benefit from at least two semesters of ELI writing instruction.*	*Paper shows evidence of:* • Undeveloped or unclear argument • Simple topic description/ restatement, but with little insight • A general lack of supporting evidence, detail, examples • Redundancy of ideas, argumentation	*Paper is:* • Not cohesive • Formulaic (e.g., 5-para essay format), or lacking organization • Marked by absence of clear transitions between ideas, or simple sentence-level transitions used at paragraph level (e.g., first, next, then)	*Paper has:* • Notably limited vocabulary • Repetition/overuse of certain lexical items • Numerous problems with word choice • Incorrect collocations	*Paper has:* • Numerous errors that typically interfere with comprehension • General lack of sentence complexity	*Amount of writing is:* • Unsuitable for level of analysis and/or amount of time provided to write paper

Appendix B: New ELIPT Writing Test Rubric

	Content (5-10)	**Organization (5-10)**	**Vocabulary (5-10)**	**Grammatical Complexity (5-10)**
9-10 *Shows high proficiency in L2, knowledge and control of academic writing genres/ conventions common in US universities.*	*Paper shows evidence of:* • Clear, highly developed point/argument • Complexity of thought/ analysis • Insight on the topic, rather than mere description • Ample supporting evidence, detail, and examples • Command of rhetoric, evidence of writing style	*Paper is:* • Cohesive • Well-developed • Unique (not formulaic) • Marked by clear transitions, appropriate use of transitional phrases	*Paper has:* • Wide variety of vocabulary • Appropriate use of idioms • Few problems with collocations • Few problems with word choice	*Paper has:* • Complex sentence structure (e.g., complex coordination, subordination, embedded questions, etc.) • Some errors, but none that interfere with comprehension
7-8 *Shows some knowledge and control of academic writing; needs to develop L2 proficiency, writing ability, and/or awareness of genres/conventions common in US universities.*	*Paper shows evidence of:* • Clear, developed argument, but may be simplistic • Some insight on the topic, but may lack depth • Effective support, but evidence and examples may be general or vague	*Paper is*: • Cohesive • Somewhat formulaic • Marked by appropriate transitions, with some misuse/overuse of transitional phrases	*Paper has:* • Varied vocabulary • Some problems with collocations • Some problems with word choice	*Paper has:* • Some correct complex sentence structure; evidence of other (incorrect) attempts • Several errors (e.g., tense/aspect, word form, articles, prepositions), but typically do not interfere with comprehension
5-6 *Needs to develop L2 proficiency; notable unfamiliarity with and general lack of control of academic writing; would benefit from at least two semesters of ELI writing instruction.*	*Paper shows evidence of:* • Undeveloped or unclear argument • Simple topic description/ restatement, but with little insight • A general lack of supporting evidence, detail, and examples • Redundancy of ideas, argumentation • Not enough to evaluate	*Paper is:* • Not cohesive • Formulaic, or lacking organization • Marked by absence of clear transitions between ideas, or simple sentence-level transitions used at paragraph level (e.g., first, next, then) • Not enough to evaluate	*Paper has:* • Notably limited vocabulary • Repetition/overuse of certain lexical items • Numerous problems with word choice • Incorrect collocations	*Paper has:* • General lack of sentence complexity • Numerous errors that typically interfere with comprehension

Explanatory Note: Because students in the ELI are at a level of academic English proficiency that is high enough to be admitted to UH Manoa, the 6-point numbering system used here ranges from 5-10, rather than 1-6.